PERFORMING DATA LINEAGE FOR AN INGESTION SYSTEM TO A DATA LAKE

MEMÒRIA TÈCNICA DEL PROJECTE

1 D'ABRIL DE 2016

Autor: Marc Garnica Caparrós
Director: Albert Abelló
Ponent: Pedro González

Especialitat: Enginyeria del Software
Facultat d'Informàtica de Barcelona, UPC

RESUM

Aquest treball de final de grau engloba el disseny, desenvolupament i documentació d'un dels components del sistema WISCC. El World Information System for Chagas Control és un projecte propulsat per l'Organització Mundial de la Salut i actualment en desenvolupament pel grup Database Technologies and Information Management de la Facultat d'Informàtica de Barcelona.

El projecte WISCC té el principal objectiu de construir un magatzem de dades provinents de nombroses i disperses fonts, tant en estructura com en contingut, relacionades amb la malaltia del Chagas. Aquest treball de final de grau documenta el desenvolupament complet del connector entre el sistema WISCC i el seu mòdul principal d'entrada de dades per part dels usuaris, implementat i distribuït mitjançant l'eina software DHIS2 . Aquest treball engloba des dels processos d'extracció per obtenir les dades de la font externa fins a la ingestió de les dades en el repositori central WISCC, mantenint un control i una anotació del flux de dades generat.

ABSTRACT

This bachelor degree project includes the design, development and documentation for a subcomponent of the WISCC System. The World Information System for Chagas Control is an ambitious project launched by the World Health Organization and currently being developed by Database Technologies and Information Management group in the Facultat d'Informàtica de Barcelona.

WISCC project has the main goal to build a data repository integrating data from disparate sources related with Chagas disease. This project contains the complete development of the connector between the main WISCC System and its main manual data source, implemented and distributed through DHIS2 software tool. The documentation includes from the main software systems to obtain and extract all the information introduced on the external tool, to the main software systems to ingest the new data to the repository, maintaining annotations and control of this data lineage.

Aprofito l'ocasió per agrair a Oscar Romero i Albert Abelló l'oportunitat de participar en aquest projecte.

Gràcies a Pedro González i Eric Mourín per la seva ajuda.

I finalment també un agraïment a la Judit i a la meva família per tot el recolzament i ajut.

ÍNDEX

ÍNDEX D'IL·LUSTRACIONS 2

ÍNDEX DE TAULES 3

1. INTRODUCCIÓ 4

1.1 ... CONTEXT 4

1.2 ... ESTAT DE L'ART 9

1.3 ... FORMULACIÓ DEL PROBLEMA 11

1.4 ... ABAST DEL PROJECTE 12

1.5 ... METODOLOGIA DE TREBALL 14

2. PLANIFICACIÓ TEMPORAL 15

2.1 ... CALENDARI I VISIÓ GLOBAL 15

2.2 ... ROLS EN EL DESENVOLUPAMENT DEL PROJECTE 15

2.3 ... DESCRIPCIÓ DE LES TASQUES DEL PROJECTE 15

2.4 ... SPRINTS – ITERACIONS DE DESENVOLUPAMENT 16

2.5 ... VALORACIÓ D'ALTERNATIVES I PLANS D'ACCIÓ 18

2.6 ... DESVIACIONS DE LA PLANIFICACIÓ 20

3. GESTIÓ ECONÒMICA 21

3.1 ... IDENTIFICACIÓ I ESTIMACIÓ DELS COSTOS 21

3.2 ... CONTROL DE GESTIÓ 24

4. DISSENY DEL SISTEMA 26

4.1.... SISTEMA D'EXTRACCIÓ 26

4.2.... SISTEMA D'INGESTIÓ DE LES DADES 29

5. DESENVOLUPAMENT DEL SISTEMA 34

5.1.... DESENVOLUPAMENT DEL SISTEMA D'EXTRACCIÓ 34

5.2.... DESENVOLUPAMENT DEL SISTEMA D'INGESTIÓ 39

6. INFORME DE SOSTENIBILITAT 41

6.1.... DIMENSIÓ ECONÒMICA....41

6.2.... DIMENSIÓ SOCIAL....41

6.3.... DIMENSIÓ AMBIENTAL....42

7. CONCLUSIONS 43

7.1 ... ASSOLIMENT DELS OBJECTIUS....43

7.2 ... CONEIXEMENTS APLICATS I ADQUIRITS 43

7.3 ... TREBALL FUTUR 44

8. REFERÈNCIES 45

ANNEX A: DIAGRAMA DE GANTT....46

ANNEX B: DIAGRAMA DE GANTT de DESPRÉS DE LES DESVIACIONS 53

ÍNDEX D'IL·LUSTRACIONS

Figura 1 WISCC system overview 6

Figura 2 Arquitectura de desplegament del sistema WISCC diferenciant les instàncies de DHIS2 8

Figura 3 Abast del projecte 13

Figura 4 Disseny sistema d'extracció 26

Figura 5 Flux del procés d'extracció 27

Figura 6 Graf d'estats de les dades en el sistema WISCC 28

Figura 7 Esquema general del sistema d'ingestió de les dades 29

Figura 8 Funcionament de la gestió del canal DHIS2 mitjançant les ontologies 31

Figura 9 Execució DHIS2-CONNECTOR en la instància global 37

Figura 10 Sistema d'extracció per instàncies locals 38

Figura 11 Visió global del projecte 46

Figura 12 Anàlisi previ i introducció: Inception 47

Figura 13 Disseny i implementació del sistema d'extracció de dades: Sprint 2 48

Figura 14 Disseny i implementació del sistema d'extracció de dades: Sprint 3 49

Figura 15 Disseny i implementació del sistema d'injecció de dades al WISCC: Sprint 4 50

Figura 16 Disseny i implementació del sistema d'injecció de dades al WISCC: Sprint 5 51

Figura 17 Avaluació final: Sprint 6 52

Figura 18 Desviacions de la planificació Sprint 4 53

Figura 19 Desviacions de la planificació Sprint 5 54

Figura 20 Desviacions de la planificació Sprint 6 55

Figura 21 Desviacions de la planificació Sprint 7 56

ÍNDEX DE TAULES

Taula 1 Valoració d'alternatives i plans d'acció 19

Taula 2 Costos de recursos humans desglossats per activitats de Gantt 22

Taula 3 Costos de software i llicències 22

Taula 4 Costos generals del projecte 23

Taula 5 Informe de costos totals del projecte 24

Taula 6 Càlcul de desviacions per tasca del projecte 25

Taula 7 Matriu de Sostenibilitat del projecte 41

1. INTRODUCCIÓ

1.1 CONTEXT

1.1.1 ORGANITZACIÓ MUNDIAL DE LA SALUT I LES MALALTIES TROPICALS DESATESES

La Organització Mundial de la Salut (OMS) és una organització constituent de les Nacions Unides creada l'any 1948 [1]. Més enllà d'especialistes en Salut, científics i epidemiològics, la corporació de la OMS compta amb persones especialitzades en la gestió administrativa, finances i de sistemes de la informació; com també amb experts en economia i estadística aplicada a la salut.

L'Octubre de 2010 la OMS va publicar el primer informe sobre les Neglected Tropical Diseases (NTD)[2], malalties tropicals desateses. Un subconjunt de malalties infeccioses que formen un grup fortament associat amb la pobresa. La OMS va estimar que aproximadament un bilió de persones provinents de 149 països endèmics estaven afectades per almenys una d'aquestes 17 malalties tropicals.

El gener de 2012, la OMS va publicar un full de ruta a seguir, fixant els principals objectius de prevenció, control, eliminació i eradicació de les 17 NTD i el gener de 2013 va publicar el segon informe[3], el qual definia els conceptes primordials de control, eliminació i eradicació; analitzava els principals reptes a nivell de país, emfatitzava la millora en la coordinació i integració a nivell de país, ressaltava l'enfortiment dels recursos humans i recalcava la necessitat de treballar en altres sectors importants com són l'educació o l'agricultura, entre d'altres.

Aquest seguit d'iniciatives i publicacions estan actualment en procés i demostren visibles resultats en termes de disponibilitat de la informació sobre les NTD, renom en els mitjans de comunicació i augments considerables en la consciència del problema en els sistemes de salut, comunitats científiques i organitzacions no governamentals. Produint una conseqüent onada de respostes com donacions de medicaments, fons monetaris i noves iniciatives i publicacions per ressaltar la importància de les NTD.

1.1.1 CHAGAS DISEASE

Chagas Disease és una de les 17 malalties desateses incloses dins de les NTD. També coneguda com la Human America Trypanosomiasis [3] és una malaltia potencialment mortal causada pel paràsit protozoari Trypanosoma cruzi (T.cruzi). Es troba principalment en les zones endèmiques de 21 països de l'Amèrica Llatina, on és majoritàriament transmesa per vectors als humans per mitjà del contacte amb la femta del triatomine bug (kissing bugs), el principal insecte portador i transmissor.

Durant el tercer comitè d'Iniciativa dels Països no Endèmics convocat per la Organització Mundial de la Salut, el juny de 2013, es va presentar l'anomenada Tricycle Strategy of the Progamme on Control of Chagas disease, promoguda pel departament de la OMS del control de les NTD. Aquesta estratègia consisteix bàsicament en dos motors d' actuació: interrupció de les transmissions i la prestació d'atenció a la població afectada; i un volant d'actuació: un **sistema d'informació responsable de l' aprovisionament d'informació i vigilància**. La construcció d'aquest sistema té un important

valor addicional, crear i promoure consciència sobre la malaltia del Chagas, especialment facilitant l'accés a dades interactives, estadístiques d'afectats, mapes i diagrames il·lustratius.

1.1.2 WISCC: WORLD INFORMATION SYSTEM FOR CHAGAS CONTROL

Els projecte WISCC està actualment en producció per l'equip d'investigació i recerca Database Technologies and Information Management group (DTIM) del departament de Serveis i Sistemes de la Informació, a la Universitat Politècnica de Catalunya.

L'objectiu del projecte WISCC és implementar un sistema d'informació pel control de la malaltia del Chagas. El principal objectiu tècnic del projecte és la construcció d'un repositori de dades integrat amb dades de fonts molt disperses en quant a freqüència, format, semàntica, naturalesa i estructura però amb la malaltia com a clar lligam multidisciplinari.

Moltes relacions potencials de dades sovint no es coneixen per endavant, l'estructura de les dades i els esquemes a seguir poden canviar, de la mateixa manera que els patrons d'anàlisi d'aquestes dades i el seu consum principal han d'evolucionar. Dins d'aquest escenari, la capa base arquitectònica del repositori de dades a construir adoptarà la forma d'un Data Lake, un magatzem de dades sense estructura prèvia on totes les dades es guarden amb el seu format original. Sobre aquesta capa base, noves capes es poden construir d'una manera flexible i dinàmica. Aquestes capes s'alinearan amb les necessitats específiques sol·licitades pels usuaris analítics que no són predibles ja que són canviants en funció dels requisits del Programa contra les malalties desateses de la OMS. Per aquesta raó el projecte a desenvolupar està destinat a ser una eina valuosa, tant a curt com a llarg termini.

Aquest projecte es va iniciar l'any 2015 i tots els seus requisits van ser àmpliament documentats, així com l'Estat de l'Art del sistema i l'anàlisi i disseny software conseqüent per Jaume Viñas en el seu Projecte de Final de Grau[6].

1.1.2.1 SYSTEM OVERVIEW

La figura 1 conté l'esquema principal del Sistema WISCC complet. Consistirà d'un repositori central de dades sobre el Chagas governat per un gestor de metadades i tindrà adjuntat tant el mòdul d'entrada de dades per part dels usuaris, com el mòdul d'anàlisi de dades i les diferents interfícies d'interacció amb el sistema.

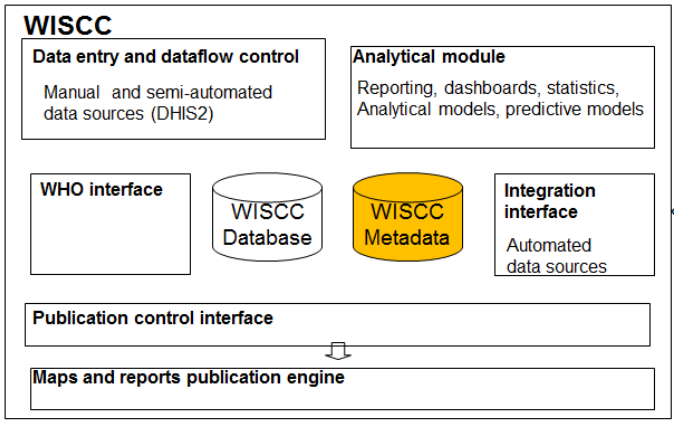

Figura 1 WISCC system overview

El sistema tindrà quatre principals interfícies d'interacció amb diferents tipus i orígens d'usuaris. En aquest punt és convenient definir clarament quins van ser els actors que es van concretar pel projecte WISCC[6]:

- La **OMS** en sí, formaria part del conjunt d'actors del sistema ja que interactuarà amb ell per visualitzar, gestionar, analitzar i importar/exportar tota mena d'informació, com també tindrà accés a tota la informació emmagatzemada en el repositori: consultar (sense modificar) tota la informació afegida per qualsevol país. Addicionalment la OMS també serà responsable de gestionar les avaluacions programades, regulacions i certificacions.
- El sistema d'informació tindrà la seva font principal de dades provinent dels actors: **Health minister officers(HM)** and **Researchers(R).** Tants els investigadors com els treballadors dels ministeris de salut dels diferents països seran responsables de proveir, modificar i gestionar informació mèdica d'interès del seu propi país o organització no governamental. També seran capaços de visualitzar la seva informació mitjançant mapes i diagrames. Aquests actors interactuaran directament amb el mòdul Data Entry and Data Flow, mòdul primordial pel desenvolupament d'aquest projecte.
- El sistema WISCC també preveu l'adherència d'actors representats per altres sistemes d'informació. Mitjançant la interfície d'integració el sistema serà capaç d'integrar informació provinent d'altres sistemes com ProMED, PubMed o Google Alerts.

1.1.3 DATA ENTRY AND DATA FLOW CONTROL

El mòdul Data Entry del sistema és un dels mòduls més essencials i crítics ja que representa la interfície d'entrada de dades més utilitzada i nombrosa de tot el sistema. Mitjançant aquest mòdul la informació dels usuaris (actors HM i R) s'integrarà dins del repositori. El ministeris de salut i els centres d'investigació entraran aquesta informació per mitjà de formularis mèdics i administratius. El disseny i implementació d'aquests formularis és una de les parts més importants doncs de la creació del mòdul Data Entry. Aquets formularis seran la entrada principal de dades mèdiques que posteriorment dins del sistema WISCC seran objecte d'anàlisis exhaustiu. La interfície dels formularis requereix ser amable amb l'usuari i còmode d'usar.

DTIM grup va explorar diferents possibilitats per aconseguir dissenyar i implementar aquests formularis de forma eficient, tenint en compte que tot i la seva importància, la part clau del sistema WISCC no serien el seu data Entry sinó els seus mòduls d'anàlisis i integració de metadades. És per això que entre les opcions de crear i implementar des de zero una interfície còmode, accessible i usable per l'usuari i aprofitar eines open-source distribuïdes al mercat per la gestió de la Entrada de dades es van decantar per la segona opció.

1.1.3.1 DHIS2 TOOL

Dhis2 és bàsicament una eina web modular mitjançant Java Frameworks desenvolupada pel Health Information System Programme (HISP) com una eina oberta a la comunitat científica i àmpliament distribuïda. És una eina molt usada arreu del món, genèrica, sense cap pena de pre-configuració que permet gestionar el model de metadades de forma flexible i permet a l'usuari dissenyar els continguts pel seu sistema d'informació sense cap mena d'implementació o programació.

Finalment, les pautes definides en la implantació de l' eina DHIS2 com a sistema software responsable de l' entrada de dades del sistema WISCC van definir clarament que l' eina simplement cobriria les funcionalitats de generació de formularis i validacions bàsiques i que s'utilitzaria de la forma més senzilla, sempre maximitzant la simplicitat i el desenvolupament ràpid dels formularis. Un cop correctament parametritzada l'eina, es distribuiria als països adherits al programa.

Fruit d'aquesta delimitada utilització de l'eina DHIS2 dos conceptes de disseny s'han definit dins del desenvolupament del sistema WISCC: La distribució de l' eina als diferents països adherits al programa i la importació de les dades introduïdes al DHIS2 cap al sistema central WISCC.

1.1.3.2 IMPLANTACIÓ DEL MÒDUL DATA ENTRY: ARQUITECTURES DE DESPLEGAMENT

Un cop tots els formularis es configurin correctament i l' eina estigui àmpliament confeccionada per l'ús dels actors principals del mòdul d'entrada de dades (Health minister officers i Researchers), es distribuirà als diferents països adherits al programa de control de la malaltia del Chagas. Aquests països podran accedir al mòdul d'entrada de dades del sistema mitjançant un navegador web.

La implantació del mòdul d'entrada de dades però compta amb certes diferències entre els països degut al seu grau de confidencialitat i de manteniment de les dades. S'han definit per tant dos arquitectures principals de desplegament altament associades amb els dos tipus diferents d' instàncies que s'implantaran del mòdul d'entrada de dades mitjançant DHIS2, com es pot veure il·lustrat a la figura 2.

En primer lloc es defineix la **DHIS2 LOCAL INSTANCE,** instància local de l' eina on l'aplicació web, la plataforma i la base de dades pròpia del DHIS2 estaran localitzades dins dels recursos computacionals del país en qüestió. És el cas de països amb la voluntat de mantenir i gestionar les seves dades fins donar permís explícit de que poden ser enviades al sistema central integrat WISCC. La informació parametritzada de la instància local contindrà informació solament relacionada amb el país determinat.

Per separat existirà la **DHIS2 GLOBAL INSTANCE,** una instància global de la eina on l'aplicació web, plataforma i base de dades pròpia de l'eina DHIS2 estaran localitzades i centralitzades dins les dependències de la Organització Mundial de la

Salut. Tot i que la instància contindrà informació de tots els països associats al programa de control del Chagas, cada usuari només podrà accedir a informació (dades o metadades) relacionada explícitament amb el seu país o entorn d'actuació.

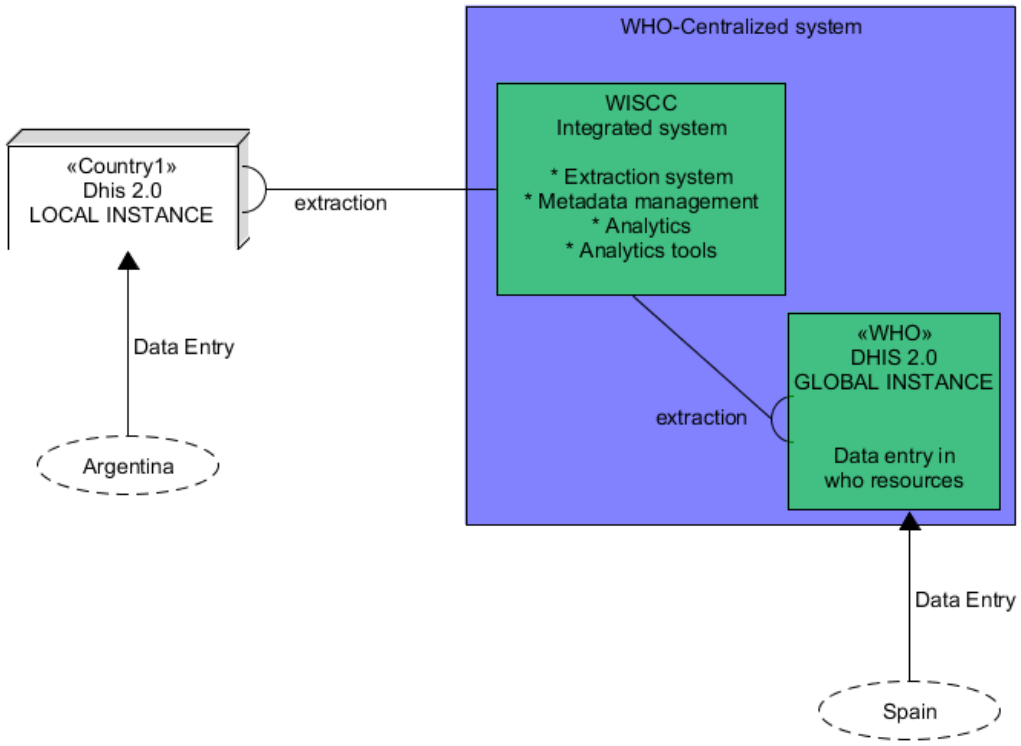

Figura 2 Arquitectura de desplegament del sistema WISCC diferenciant les instàncies de DHIS2

1.1.3.3 EXTRACCIÓ DE LES DADES I INGESTIÓ AL SISTEMA WISCC

L'aplicació DHIS2 contribueix positivament en el desenvolupament del mòdul d'entrada de dades, l'usuari introduirà en els formularis configurats la informació adient, de forma còmode i usable a través d'un portal web que interconnectarà usuari i instància de l' eina, però aquestes dades estaran localitzades dins del sistema intern de DHIS2.

Per tant es requereix d'un subsistema software que a partir de les dades introduïdes a l' eina sigui capaç d'exportar-les i injectar-les dins del sistema WISCC emmagatzemades en el repositori central. Un sistema que mantingui el flux de les dades, gestioni un historial de versions i sigui capaç d'exportar de forma eficient i controlada la informació que s'afegeix per mitjà de DHIS2 tenint en compte els requeriments de cada tipus d'instància i capaç d'emmagatzemar-ho a la primera capa del repositori del sistema WISCC.

L'objectiu principal d'aquest projecte és el disseny, anàlisis i implementació d'aquest subsistema software capaç de connectar l'eina DHIS2 amb el sistema central WISCC.

1.2 ESTAT DE L'ART

1.2.1 DATA WAREHOUSES I DATA LAKES

Tot i no ser un projecte amb amplis i clarividents factors per constituir un projecte Big Data, el sistema WISCC es construirà conforme a una de les filosofies principals del corrent Big Data i que consisteix en l' alliberació del repositori de tota mena d'estructura definida a priori que encaixi i limiti les virtuts analítiques de les dades introduïdes. El repositori central de dades WISCC respondrà en el seu nivell més bàsic amb el que les tendències més modernes anomenen un data Lake, un magatzem d'integració de dades que difereix dels ja tradicionals Data Warehouses i que afavoreix a l'evolució, enriquiment i canvi de les dades sense requerir d'una reestructuració del sistema.

El concepte de Data Lake[8] es defineix com un magatzem de dades estructurades o no. A diferencia dels Data Marts o Data Warehouses, que són magatzems de dades òptims per analitzar dades fixant-ne el seus processos d'emmagatzematge i nivells d'agregació, un data Lake està dissenyat per mantenir el format original de les dades sense pressuposar o definir a priori cap ús o abast d'aquestes dades. Els esquemes i estructures necessaris en les dades introduïdes es van configurant a mesura que aquestes dades són utilitzades, aquest concepte de creació de l'esquema en temps de consulta és l'anomenat **late binding** o **read schema[8]**.

L'aspecte més innovador d'aquest nou concepte és el fet que el sistema estalvia molt de temps utilitzat en la preparació de les dades, com se sol prioritzar en data Warehouses. Emmagatzemant les dades amb el seu format original, sense passar per processos de modulació, deslliga el conjunt sencer de dades d'un únic model conceptual. [9]El concepte de data Lake no només és útil en termes de repositoris massius de dades si no també en quan a generar un nou ecosistema de dades. Totes les dades estan disponibles en un data Lake, sense cap mena de restricció. Des d'un punt de vista analític podem generar en qualsevol moment interseccions i creuaments entre elles i generar ràpidament altres formes de coneixement.

El model data Lake és actualment usat per companyies com Google, Bing o Yahoo per emmagatzemar grans quantitats de dades molt diferents. La tecnologia que suporta el concepte de data Lake és Hadoop. L'arquitectura és molt simple: un Sistema de Fitxers Hadoop organitzat mitjançant directoris i fitxers.

El punt clau del disseny i implementació d'un data Lake és el concepte de metadada (dada sobre una dada). Totes les dades entrades al sistema (podríem definir-les com dades injectades), estan associades amb metadades que són les responsables de identificar-les, localitzar-les i catalogar-les. Un sistema data Lake practica el concepte de metadada mitjançant múltiples etiquetes (tags): Un certa dada pot ser fàcilment localitzable i compresa a partir d'un cert conjunt de tags.

1.2.2 PROCESSOS ETL

Els processos ETL (Extraction, Transformation and Load) [10] són processos àmpliament utilitzats en Data Warehouses i Data centers responsables de capturar les dades dels diferents orígens disponibles i introduir-les en el magatzem de dades. Un procés ETL conté aquestes parts principals:

EXTRACCIÓ [11] de les dades provinents d'altres sistemes d'informació, dades que convergeixen en un definit i consolidat format de dades del Data Warehouse preparat pels processos següents. Existeixen dos tipus primaris d'extracció:

- Full extraction: Les dades disponibles són completament extretes del sistema d'origen. Com que transportem totes les dades no hi ha necessitat de mantenir un registre de canvis des de l'última extracció, no es requereix de cap altra informació (com podrien ser Timestamps).
- Incremental extraction: En un cert moment, es comparen els estats de la base de dades interna del sistema amb la del sistema extern i s'extreuen tots els canvis detectats que conseqüentment s'importen al sistema central. Per identificar les dades que s'han d'extreure es requereix d'atributs addicionals en les dades que reflecteixin la seva última modificació o versió. Molts data Warehouse no realitzen camp mena de computació addicional per realitzar una extracció incremental, simplement realitzen una extracció completa i comparen amb l'estat del seu magatzem de dades i l'ultima extracció per identificar les dades noves o modificades.

TRANSFORMACIÓ [12] de les dades on a partir de les dades extretes s'apliquen regles de negoci, derivacions, agregacions, dimensions, neteges, filtres, unions, encreuaments i tota mena d'operacions de validació o comprovació de la correcció de les dades.

Un concepte molt important a definir és l'anomenat **Transformation Flow**:

- Multistage data transformation: La lògica de transformació conté múltiples passos que finalitzen en un estat vàlid per la càrrega.
- Pipelined data transformation: Solapament dels processos de càrrega i transformació per millorar el flux de dades i augmentar l'escalabilitat del sistema.

CÀRREGA [12] de les dades en el magatzem o Data Warehouse per ser usat per altres aplicacions, per generar informes i panells de control.

1.2.3 CONCLUSIÓ

Els conceptes de Data Lake i de processos ETL són primordials per la concepció d'aquest projecte. El seu principal objectiu és la generació d'un sistema software que practiqui els conceptes definits en els processos ETL però que seguint amb la filosofia dels Data Lake no apliqui cap mena d'esquema previ o modelització a les dades extretes.

El procés a seguir per part de les dades hauria de ser el menys invasiu possible i emmagatzemar totes les dades provinents de les instàncies de DHIS2 amb el seu format original. Finalment els processos del sistema software a construir haurien de consumir fitxers de configuració i propietats externes per abstreure al més alt nivell la lògica de funcionament del sistema i anàlogament generar fitxers de control d'errors i depuració per poder mantenir un exhaustiu control de tasques realitzades i modificacions de les dades.

1.3 FORMULACIÓ DEL PROBLEMA

La necessitat que ha sorgit en el projecte ja ha estat contextualitzada i explicada en apartats anteriors. L'ús de l'eina DHIS2 comporta molts beneficis en temps de rapidesa i comoditat de desenvolupament i també facilitats d'ús de cara l'usuari.

Però per altra banda també comporta que les dades una vegada inserides pels usuaris mitjançant la ínterficie de DHIS2, no están localitzades dins del domini del sistema. Es per això, que es necesita un sistema software capaç d'exportar les dades afegides a les diferents instàncies de l'eina DHIS2 i les emmagatzemi en el repositori central.

Els objectius d'aquest projecte són fortament governats per les directrius de disseny que es persegueixen el projecte WISCC.

1.3.1 OBJECTIUS PRINCIPALS DEL PROJECTE WISCC

El sistema WISCC té 3 objectius o premisses imprescindibles de disseny que regeixen els objectius d'aquest projecte i que per tant és important esmentar:

1. Gestió àgil de la variació estructural de les fonts d'entrada . El projecte WISCC és un projecte de llarg termini tant en el seu desenvolupament com en el seu ús. És per això que es prioritza la seva independència de les estructures i requisits, tant dels formats d'entrada de les dades com de les peticions d'anàlisis i tractament. Es preveu que aquestes dues estructures puguin canviar molt freqüentment i es vol generar un sistema que s'adapti amb facilitat als canvis.
2. Govern per part dels usuaris de negoci del cicle complet de les dades. El sistema WISCC es cataloga com un projecte de Master Data Management[4].
3. Independència total de les eines analítiques d'explotació de les dades per part del sistema central.

1.3.2 OBJECTIUS ESPECÍFICS DEL PROJECTE

Derivant directament dels objectius 1 I 2 del projecte WISCC aquest projecte té dos objectius principals a assolir:

O1. Descodificació i tractament paramètric de la font d'informació externa. Ús totalment autònom de les metadades del sistema per l'extracció i tractament de les dades.

O2. Selecció i creació d'una BDD schema-less per emmagatzemar les dades obtingudes. Aconseguir un emmagatzemament totalment independent de cap mena de format o restriccions per tal de tenir total llibertat alhora de modelar les dades.

1.3.3 OBJECTIUS TÀCTICS DEL PROJECTE- DIRECTRIUS DE DISSENY

Un cop definits els objectius principals que guiaran i justificaran totes les accions i condicions del projecte, es defineixen aquestes principals objectius tàctics o directrius de disseny que persegueixen l'assoliment tant de O1 com o2:

OT1. Extracció de les dades incremental des de les instàncies externes de la eina DHIS2 desplegada pel projecte WISCC mitjançant una API web.

OT2. Tractament de les dades extretes per eliminar qualsevol format i emmagatzemament en cru d'aquestes dades a la primera capa del repositori central de dades.

OT3. Aprovisionament de requisits tècnics i configuracions que alimentin els processos creats, per mantenir una total parametrització dels tractaments de la font de dades.

OT4. Anotació del Data lineage i flux de dades en tot el sistema complert que generi una interfície informativa de les accions dutes a terme pel sistema total.

1.4 ABAST DEL PROJECTE

Aquest projecte engloba les fases de disseny, anàlisis i implementació d'un sistema d'extracció de les dades introduïdes a l' eina DHIS2 distribuïda als països adherits al programa de control de la malaltia del Chagas i un sistema d'integració de les dades provinents d'aquesta extracció que apliqui un format primari als formularis i els emmagatzemi directament en el data Lake del sistema WISCC on conformaran la capa base del repositori.

El desenvolupament d'aquest sistema estarà dividit en tres fases principals. La primera fase serà l'anàlisi dels protocols d'exportació pels quals han de passar els formularis introduïts a l'eina DHIS2 definint els requeriments d'usuari conforme a les validacions de les dades introduïdes i la privacitat de la informació. És a dir quins estats han de ser previstos per cada formulari per ser finalment introduït en el sistema central. En aquesta part entra molt en joc l'anàlisi dels requisits d'usuari diferenciant entre aquells països amb instància global del DHIS2 i els països amb instància local.

Seguidament, la següent fase es dividirà en dues parts importants: Per una banda s'haurà de realitzar el disseny conceptual del sistema d'exportació. En aquesta fase s'hauran de definir amb detall totes les estructures de dades necessàries per la gestió de l'exportació (quan i com s'haurà d'extreure dades de les instàncies de DHIS2) i també definir les estructures de dades addicionals que ens permetran mantenir un historial de tasques del sistema de exportació . Anàlogament, l'altra part de la segona fase serà el disseny de les estructures de dades, tant de gestió com addicionals, del sistema d'injecció de les dades extretes en el repositori del sistema WISCC. Aquesta part comportarà també un previ estudi i disseny conceptual del model de dades primari i poc intrusiu que hauran de tenir els formularis introduïts a la primera capa del data Lake.

La tercera fase consistirà en la implementació gradual dels sistemes dissenyats. Començant pels processos de governació i responsables de guiar tots els sistemes, implementar els processos funcionals que extrauran, tractaran o injectaran les dades i finalment implementar els processos responsables de les anotacions i la gestió d'un historial detallat de les tasques realitzades pel sistema, no només per poder documentar la feina realitzada sinó també per poder restablir un estat correcte del sistema en cas de fallada inesperada.

L'abast final d'aquest projecte es reflecteix directament en el diagrama de la imatge següent:

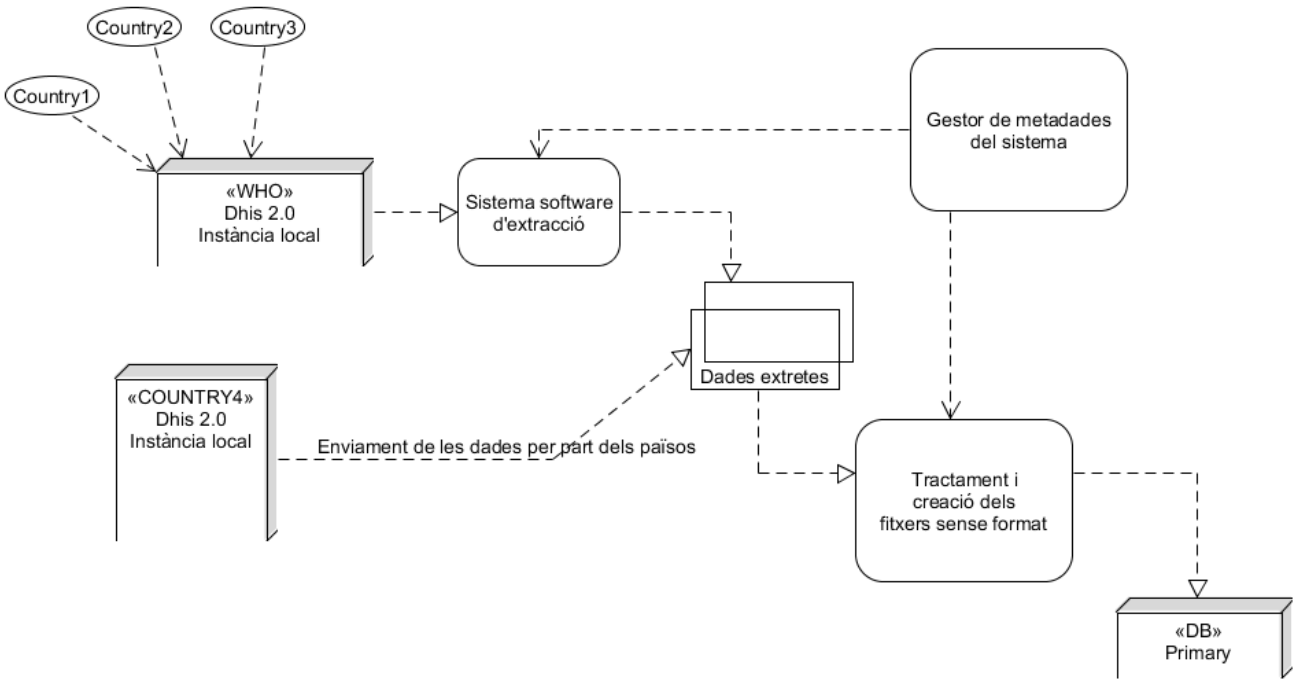

Figura 3 Abast del projecte

1.4.1 POSSIBLES OBSTACLES I PLANS DE CONTINGÈNCIA

Un dels principals obstacles a tenir en compte és l'alta dependència que presenta el projecte d'extracció amb l'eina externa DHIS2. No només perquè és una eina en desenvolupament open-source i per tant factible de contenir errors funcionals si no també perquè primer de tot es requereix d'un estudi exhaustiu de les possibles opcions a l'hora d'extreure les dades del seu sistema. Tenint en compte les funcionalitats observades actualment, l'eina DHIS2 mostra àmplies opcions a nivell d'interfície amb l'usuari i també mitjançant una API web per l'extracció de les dades introduïdes. Tot i això, també s'haurà de contemplar la possible extracció de la informació del seu sistema mitjançant processos directament cridats a la seva base de dades. Fet que ens assegura que almenys una possible opció per extreure les dades serà factible de ser implementada.

Els altres possibles obstacles formen part del conjunt d'errors de comprensió i enteniment amb els requisits del client. Ja sigui fent referència als processos de validació, privacitat de les dades, freqüència d'abstracció o gestió de versions. El millor mètode de prevenció respecte aquests riscos és clarament una constant retroalimentació amb el client mitjançant reunions freqüents i consultes.

1.5 METODOLOGIA DE TREBALL

La metodologia de treball escollida pel desenvolupament d'aquest projecte és la metodologia Àgil SCRUM que es duu a terme de forma anàloga en el desenvolupament del sistema WISCC. Aquesta metodologia aporta a l'equip una gestió regular de les expectatives i requeriments del client, resultats i avaluacions freqüents i alhora flexibilitat i alta adaptació als canvis. En conseqüència molts dels riscos es podran preveure i mitigar prèviament.

Les reunions seran molt més abundants i necessàries al principi del projecte, sobretot per discutir i definir els protocols d'extracció de les dades localitzades en les diferents instàncies de l'eina pel mòdul d'entrada de dades.

1.5.1 EINES DE DESENVOLUPAMENT

Les eines per la planificació de les tasques a realitzar i la gestió de l'equip de desenvolupament del projecte WISCC són per una banda el Microsoft Project per l'avaluació de costos temporals i assignació de tasques i per altra banda la plataforma online Trello [13], una plataforma web per gestionar projectes àgils mitjançant task boards.

Actualment les instàncies de DHIS2 s'executen en un servidor Tomcat i es desenvolupen mitjançant Java Frameworks. Durant la primera fase del projecte s'analitzaran més detingudament les possibles opcions d'implementació però es preveu la utilització de programes en Java per la creació dels processos tant d'extracció i injecció en el sistema de fitxers com la creació de fitxers d'historial de tasques i log d'errors. Per implementar la injecció de les dades en el primer nivell del repositori Data Lake existeix un clar lligam amb la tecnologia elegida per implementar aquesta capa i les seves funcionalitats d'ús.

1.5.2 MÈTODES DE VALIDACIÓ

Les validacions seran periòdiques seguint amb la filosofia de desenvolupament àgil tant pel que fa al disseny conceptual de les diferents fases com demostracions de petites funcionalitats incorporades.

Durant les reunions pròpies de l'equip es validaran prioritàriament les funcionalitats del sistema, la seva correcta implementació i sobretot es donarà molta importància a la correctesa tècnica del projecte, realitzant justificacions i argumentacions de cost, espai i complexitat.

Les reunions amb els clients serviran més per avaluar la correcció de les funcionalitats respecte a les necessitats expressades pels clients. El subsistema a desenvolupar està molt fortament lligat amb la validació de les dades introduïdes per parts dels diferents actors i sobretot amb la privacitat de les dades

2. PLANIFICACIÓ TEMPORAL

2.1 CALENDARI I VISIÓ GLOBAL

El projecte té una durada estimada de 4 mesos, des del 15 de Febrer fins aproximadament la primera setmana de Juny. (Deadline exacte 6/06/2016). Consta d'una càrrega de treball de 405 hores. La dedicació de recursos serà únicament d'una persona amb una dedicació aproximada de 30 hores setmanals durant el període de desenvolupament del projecte.

2.2 ROLS EN EL DESENVOLUPAMENT DEL PROJECTE

Tot i ser un projecte individual dins d'un projecte més gran englobant, el desenvolupament d'aquest projecte es desglossa en tasques pròpies de diferents rols dins l'àmbit de l'enginyeria del software. En aquest projecte s'han contemplat els rols que es mostren a continuació: Cap de projecte, analista, arquitecte, programador i responsable de proves.

Aquests rols es veuen immersos dins de les diferents iteracions que es presenten a continuació i la seva freqüència d'actuació es veu reflectida en el **diagrama de Gantt.**

2.3 DESCRIPCIÓ DE LES TASQUES DEL PROJECTE

El treball del projecte estarà estructurat en diferents fases. Aquestes fases seran el fil primordial de seguiment a les Iteracions SCRUM que es descriuran en el següent apartat i que confeccionaran les entregues i demostracions al client. La metodologia SCRUM dividirà les metes que a continuació es defineixen de les diferents fases en SPRINTS o iteracions on de forma dinàmica es definiran els objectius de cada iteració en el moment d'iniciar-la, es generà un producte al final de la iteració i s'avaluarà la seva correctesa. Les metes representen el Product Backlog del projecte, estats esTaulas que de forma incremental es pretenen assolir per aconseguir els objectius.

2.3.1 ANÀLISI PREVI I INTRODUCCIÓ

Aquesta fase serà l'inicial del projecte. Aquesta inclou tota al contextualització i definició del problema a resoldre mitjançant aquest projecte i també inclourà les primeres activitats introductòries al entorn de desenvolupament. Les principals metes d'aquesta fase són:

a. Definició de l'abast del projecte i Estat de l'Art.
b. Gestió del projecte (Temps i costos).
c. Anàlisi dels requisits generals del projecte.
d. Implantació, introducció i adequació a l'entorn de treball.

2.3.2 DISSENY I IMPLEMENTACIÓ DEL SISTEMA D'EXTRACCIÓ DE DADES DE DHIS2

Un cop el projecte s'ha iniciat i tot l'entorn teòric i pràctic està clarament definit la meta principal del projecte serà la propagació de les dades introduïdes al sistema DHIS2 cap al sistema central WISCC. En aquesta fase el projecte es centrarà

en els processos, estructures i metodologies necessàries per extreure les dades de la eina DHIS2. Les metes definides per aquesta fase del projecte són:

e. Anàlisi dels requisits del sistema d'extracció.
f. Anàlisi tècnic del sistema d'extracció: Estructures de control necessàries i pipeline de processos.
g. Anàlisi exhaustiu de la API Web proporcionada per DHIS2 per la extracció de dades.
h. Conceptualització del sistema d'extracció: Processos necessaris dins del pipeline d'extracció.
i. Implementació del sistema funcional de flux de dades.
j. Implementació dels fitxers i processos addicionals pel control de les execucions i dels fitxers d'anotació del historial.
k. Acoblament dels fitxers extrets des de les instàncies Locals de DHIS2

2.3.3 DISSENY I IMPLEMENTACIÓ DEL SISTEMA D'INJECCIÓ DE DADES AL SISTEMA CENTRAL

Un cop confeccionada la part funcional d'extracció de les dades, la següent fase del projecte consistirà en dissenyar i implementar el sistema software que injecti aquestes dades extretes en el repositori central de dades del sistema WISCC.

Les metes definides per aquesta fase són:

l. Disseny conceptual de la primera capa del repositori Data Lake del sistema WISCC.
m. Disseny del sistema de versions del registre de dades en el Data Lake.
n. Implementació d'una capa d'abstracció i ús de les metadades del sistema per generar els fitxers de creació
o. Implementació d'una primera versió de creació i injecció de dades a partir de la informació de les metadades
p. Implementació d'una versió final d'injecció de dades amb la gestió de versions.
q. Validació de la injecció de les dades

2.3.4 AVALUACIÓ I INSTAL·LACIÓ GLOBAL

Finalment la fase final del projecte consistirà en l'avaluació del sistema general acoblant el sistema d'extracció i el sistema d'injecció i la instal·lació global del sistema en el sistema integrat WISCC per començar a generar les primeres extraccions oficials.

2.4 SPRINTS – ITERACIONS DE DESENVOLUPAMENT

El procés de generació i desenvolupament del sistema es planifica en diferents iteracions que aniran seguint les metes definides en cada fase del projecte. Al final de cada iteració es disposa d'una demostració del producte final que serà factible de ser exposada al client o a l'equip de producció del sistema WISCC. Els sprints estaran dissenyats per comportar dues setmanes de treball (50 hores). Mitjançant les demostracions constants cada dues setmanes es pretén dur un control de l'estat del projecte, tant en la seva correctesa tècnica com en detectar possibles desviacions temporals o canvis en els requisits del client.

Mitjançant aquesta retroalimentació les iteracions es confeccionaran tenint en compte l'estat del projecte i les prioritats tant dels clients com de l'equip de producció del sistema WISCC. Al principi de cada sprint s'avaluaran els objectius a assolir que sempre estaran associats a una de les metes del projecte.

Cada iteració consisteix per tant en una primera fase de planificació breu, una fase de realització de les tasques i una última fase d'anàlisi i avaluació. Les demostracions a final de sprint reorganitzaran i alimentaran les prioritats dins del desenvolupament del sistema i seran punt de partida per confeccionar la pròxima iteració.

Totes les iteracions amb entregables es validaran mitjançant un sistema de validació automàtic que entrarà dades a les instàncies de DHIS2 i comprovarà el seu estat després d'un procés d'extracció i/o injecció (depenent de l'estat del projecte en el que ens trobem).

La planificació temporal d'aquests sprints i del projecte en total es poden veure en **l'Annex A: Diagrama de Gantt**

INCEPTION

Com a primera iteració es contempla una primera fase d'adequació, inception, aquesta iteració tindrà excepcionalment una durada de 4 setmanes ja que consta de tota la planificació inicial del projecte i introducció a l'entorn de treball. Aquesta sprint parteix dels objectius principals de definir i abastar tots els objectius i metes del projecte i realitzar una documentació detallada de tota la seva gestió.

D'altra banda, també té com a objectiu introduir l'entorn de treball i configuracions específiques.

SPRINT 2

La segona iteració es centra ja en la fase de disseny i implementació del sistema software de propagació de les dades de DHIS2 cap al sistema central. En aquest sprint l'objectiu principal és realitzar un estudi conceptual i tècnic dels requeriments del sistema d'extracció, valorar les possibilitats i revisar i documentar les funcionalitats que la API web de DHIS2 pot aportar.

Finalment entregar un disseny complet del sistema d'extracció i implementar una primera versió de sistema d'extracció on totes les dades de la instància global de DHIS2 es propaguin amb una determinada configuració.

SPRINT 3

Aquesta tercera iteració té com objectiu essencial ampliar la primera versió entregada en la iteració 2 amb els fitxers de configuració pertinents per poder abstraure i modificar les configuracions i així parametritzar la extracció de les dades. D'aquesta manera al final de la iteració obtenir un sistema totalment guiat a partir de fitxers de configuració on els processos funcionals de flux de dades s'alimentin de fitxers de configuració per saber què, quan i on han d'extreure les dades i actualitzin una sèrie de fitxers o estructures d'anotació de tasques realitzades i historial del sistema.

Finalment aquesta iteració també contempla l'acoblament dels fitxers extrets de les instancies Locals. Sota la premissa que es desenvoluparà dins del projecte WISCC un software de validació d'enviaments de les dades en les instàncies locals de la eina DHIS2 i que aquests enviaments estaran configurats amb el mateix format que en el flux de dades en l'extracció a la instancia Global. Aquest desenvolupament tot i formar part del projecte WISCC no es contempla dins de l'abast d'aquest projecte.

SPRINT 4

En la quarta iteració el projecte encara el desenvolupament del sistema software d'injecció de dades al sistema central WISCC. En aquest sprint l'objectiu principal seria el disseny de la primera capa del repositori central. Com a objectiu tècnic es concreta la implantació d'una capa d'abstracció de les metadades del sistema essencial per la creació dels fitxers a injectar en el repostori i es fa una primera versió del sistema d'injecció. El desenvolupament d'aquest software gestor de metadades està fora de l'abast d'aquest projecte i només es contempla com a una font d'informació per generar els fitxers a injectar en el repositori.

SPRINT 5

La cinquena iteració principalment complementarà l'entrega del sprint 4. L'objectiu principal es connectar la creació de fitxers a partir de les dades extretes amb les pautes i definicions de les metadades del sistema i generar un bon manteniment de les versions dins de la primera capa del data Lake. En aquesta iteració serà molt importar assegurar i validar el bon funcionament del sistema d'extracció i la gestió de les versions.

SPRINT 6

Finalment l'última iteració tindrà l'objectiu principal de acoblar el sistema d'extracció amb el sistema d'injecció de forma correcta, i refinar el sistema global a partir del feedback obtingut a la iteració anterior. Finalment posar en marxa de forma global el sistema d'extracció i injecció.

2.5 VALORACIÓ D'ALTERNATIVES I PLANS D'ACCIÓ

Ja que el projecte té com a premissa essencial una constant retroalimentació amb els requisits de l'equip de producció del sistema WISCC i els clients, es contempla que de forma molt probable sorgeixin desviacions temporals a les iteracions del desenvolupament del projecte. Es pot apreciar que la planificació del projecte per la seva fi contempla un cert marge en consideració amb el deadline definit pel 6 de Juny de 2016.

Durant el procés de desenvolupament del projecte es poden produir desviacions temporals a les iteracions del desenvolupament. A continuació es mostren els possibles obstacles que el desenvolupament d'aquest projecte contempla i les possibles accions al respecte:

RISC	GRAVETAT	PLÀ DE CONTINGÈNCIA
Poca experiència amb les tecnologies a utilitzar, sobretot amb els processos Script en Java a desenvolupar per l'extracció automàtica	Mitja	Avui en dia els professionals tècnics estan ensenyats per poder ser camaleònics amb les tecnologies i poder tenir (amb un temps de adequació) fluïdesa amb qualsevol d'elles. S'utilitzaran fonts d'informació contrastades i hores extres de introducció a les metodologies.
Canvis en els requisits i aparició de noves funcionalitats	Mitja	Les reunions exhaustives a cada iteració apaivagaran el pes dels canvis en els requisits i s'intentarà tenir suficient flexibilitat als canvis.
Planificació errònia	Baixa	Es mantindran reunions amb els directors i desenvolupadors del projecte WISCC per assegurar el bon desenvolupament del sistema.
Dependència de la API web de DHIS2 i de les versions de la eina. Si les versions de DHIS2 modifiquen l'estat de les dades i el format de les invocacions via API el desenvolupament es pot veure afectat.	Alta	Abstracció total de la versió de DHIS2 dins del sistema d'extracció i injecció per així tenir facilitat d'ús. En cas de canvi en la versió, identificar els canvis rellevants i canviar el fitxer de configuració corresponent. Per altra banda si la API no és suficient també es contempla poder extreure les dades directament de les bases de dades de la eina mitjançant scripts SQL.
El gestor de metadades serà qui regirà tota la generació tant de les invocacions a la API per la extracció com en la creació dels fitxers a injectar. El seu desenvolupament serà paral·lel al del sistema d'extracció i pot ser que manqui de funcionalitats o vagi endarrerit.	Alta	Activa i contínua comunicació amb els desenvolupadors del gestor de metadades per concretar les funcionalitats desitjades i les dates de finalització.
La instància local de DHIS2 necessita un software addicional de validació de les dades a enviar al sistema central que està fora del desenvolupament d'aquest sistema i que per tant pot ser que pateixi de retards	Mitja	De forma anàloga amb altres riscos es mantindrà una activa comunicació amb els desenvolupadors d'aquest software addicionals dins l'eina de DHIS2 per establir les funcionalitats i dates de finalització.

Taula 1 Valoració d'alternatives i plans d'acció

2.6 DESVIACIONS DE LA PLANIFICACIÓ

Durant el desenvolupament del projecte s'han produït fets interns i externs al projecte que han provocat canvis en el decurs del mateix, per aquest fet la planificació inicial s'ha vist afectada. Aquests són els canvis més rellevants que s'han produït:

- Uns dels primers imprevistos sorgits durant el desenvolupament del sistema va ser la llarga presa de decisió sobre les diferències en el tractament de la extracció entre la instància global de l'eina DHIS2 i les instàncies locals. En un primer moment es va dissenyar i desenvolupar un procés Java que hauria de córrer tant en els servidors locals com en el global però aquesta opció es va desestimar ja que es precisava d'accions de l'usuari per confirmar l'extracció en les instancies locals. Una vegada redireccionat el desenvolupament del projecte, la configuració de les extraccions globals es va dissenyar com una transacció amb l'usuari on era el mateix usuari qui demanava els fitxers extrets i ell era el responsable de enviar-los als servidors globals. Després de reunions amb el client es va desestimar també aquesta opció i es va acabar decidint el disseny que està presentat en aquest document, una única acció i la extracció i transferència es realitzen correlativament.
- Un dels altres punts que va provocar desviacions a la planificació va ser l'aparició de canvis importants a la API de l'eina DHIS2 després d'una actualització des de la versió 2.22 a la versió 2.23. En aquest cas, a la versió 2.22 la API permetia aconseguir tots els events actualitzats des de una certa data i en estat COMPLERT, sense cap mena de restricció de país, programa, event, etc. A la versió 2.23 els responsables de l'eina DHIS2 varen treure aquesta opció ja que generava cargues masses grans en el servidor en el cas que el volum de dades fos extremadament gran. Així doncs varen posar com a obligatori haver d'especificar o bé un país o bé un programa de formularis en qüestió a l'hora de fer l'extracció. Aquest canvi fa suposar canviar el disseny i la implementació d'una part força rellevant del projecte ja que en comptes de una sola crida a la API per aconseguir el fitxer de totes les dades extretes, el sistema ara havia d'extreure per cada programa tots els events que complien les condicions d'extracció i agrupar-los en un sol document.
- El disseny del sistema d'ingestió també va afectar a la planificació que s'havia realitzat inicialment ja que es va requerir d'un procés previ d'adequació i estudi de l'ontologia que formaria part de la descripció de les metadades, concepte molt important per dissenyar l'algorisme d'injecció el més independent de l'estructura de la font possible. Aquesta Ontologia encara està en desenvolupament per altres components del projecte WISCC i per tant el sistema d'ingestió es desenvolupa conforme l'evolució d'aquesta.

Els problemes sorgits es van produir dins de la mateixa etapa de projecte i el que van provocar és que el sprint 4 es dediqués totalment a resoldre els problemes del sistema d'extracció i a verificar el sistema d'injecció. Degut això es van haver de replantejar els sprint 4 5 i 6 conforme a les necessitats del projecte.

La planificació realitzada a partir d'aquestes desviacions s'explica en detall a l'annex B.

3. GESTIÓ ECONÒMICA

3.1 IDENTIFICACIÓ I ESTIMACIÓ DELS COSTOS

En aquest apartat del projecte es realitza una identificació i estimació dels elements claus del pressupost considerant els recursos que es proposen utilitzar. Aquests recursos es poden dividir en recursos humans (treballadors), software, hardware, costos generals i costos imprevistos. Finalment es presenta un resum amb el pressupost estimat del projecte.

3.1.1 RECURSOS HUMANS

Aquest projecte forma part d'un projecte molt més ampli i serà desenvolupat per una sola persona sota un contracte de Beca d'Aprenentatge per part del Departament de Sistemes i Serveis Software de la UPC. Aquesta persona manté un contracte laboral de 4 hores diàries, 20 hores setmanals, 80 hores mensuals amb el qual rep un sou de 470 € però que representa amb les retencions d'IRPF un cost de 490.96 €. Tot i això ja que el projecte també forma part del seu projecte de final de carrera, les hores diàries de treball estimades que es dedicaran al projecte són 5 hores diàries, 5 dies a la setmana.

Per tant, a nivell mensual es treballaran 25 x 4 = 100 hores mensuals amb un cost del personal de 490.96 €, el que resulta en un cost per hora de 4.9 €.

Amb aquesta premissa es desglossen les activitats del projecte i els seus costos a la taula 1, com es pot apreciar les tasques internes de cada iteració no s'han computat directament en el cost ja que molt sovint es desenvoluparan de forma paral·lela i contínua dins la iteració.

Fase del projecte	Tasca	Dies	Hores laborals	Cost per hora	Cost
Anàlisi previ i introducció	Inception	21	105	4.9 €	514.5 €
	Total	21	105	4.9 €	514.5 €
Disseny i extracció de dades	Sprint 2	11	55	4.9 €	269.5 €
	Sprint 3	11	55	4.9 €	269.5 €
	Total	21	105	4.9 €	514.5 €
Disseny i implementació del sistema d'injecció de dades al WISCC	Sprint 4	11	55	4.9 €	269.5 €
	Sprint 5	11	55	4.9 €	269.5 €
	Total	21	105	4.9 €	514.5 €
Avaluació final	Sprint 6	11	55	4.9 €	269.5 €
	Total	11	55	4.9 €	269.5 €
Projecte sencer		81	405	4.9 €	1984.5 €

Taula 2 Costos de recursos humans desglossats per activitats de Gantt

3.1.2 SOFTWARE

Per la gestió i el desenvolupament del projecte són necessàries eines software, tot i això la majoria d'elles comunes i gratuïtes. Una de les premisses del projecte WISCC en desenvolupar el projecte és l'ús d'eines de distribució Open-source i gratuïtes.

Producte	Cost	Vida útil	Amortització en 5 mesos
Windows 10	0 €	-	0 €
Microsoft Office 2016	0 € (descàrrega per estudiants)	4 anys	0 €
Eclipse	0 €	-	0 €
Google Chrome	0 €	-	0 €
Google drive	0 €	-	0 €
Subversion	0 €	-	0 €
MongoDB	0 €	-	0 €
DHIS2 tool	0€	-	0 €
Total			0 €

Taula 3 Costos de software i llicències

3.1.3 HARDWARE

En el desenvolupament del projecte s'usaran també recursos hardware. Per una part es tenen en compte els recursos hardware de desenvolupament i per altre els recursos de gestió i desplegament. Una vegada finalitzat el projecte el client serà el responsable de mantenir i gestionar els servidors que acolliran el sistema WISCC i per tant el sistema software resultant d'aquest projecte, ja sigui de forma directa amb Data Centers de la Organització Mundial de la Salut com amb l'ajuda de proveïdors tercers.

Tot i això, durant el procés de desenvolupament del projecte tota la gestió es portarà a terme dins dels servidors de la UPC. Aquests servidors compten amb un servei de Cloud Privat que segons les tarifes de preus de 2016 surt a 40,60 €. Com que el projecte té una durada d'uns 4 mesos assumint uns costos de hardware de **162.4 €** que ja inclou el servei de màquines virtuals i IPs públiques.

3.1.4 COSTOS GENERALS

Dins d'aquesta secció, s'inclouen els costos relacionats amb la realització del projecte, energia elèctrica consumida, internet, sol·licituds de llicencies i transport per les reunions i presentacions. El preu del kwh s'ha extret de les dades publicades per la web d'IBERDROLA l'any 2015. S'ha assumit un consum de 60 watts per part de l'ordinador de desenvolupament. L'accés a internet s'ha calculat mitjançant la mitjana de preus oferts per les companyies principals i el lloguer Cloud UPC s'ha extret directament de les seves tarifes a: **https://upcnet.upc.edu/acl_users/credentials_cookie_auth/require_login?came_from=https%3A//upcnet.upc.edu/serveis/servidors-i-xarxes/servidors/cloud-privat-upc/copy_of_tarifes-de-preus-2014**

Producte/Servei	Cost	Període	Total estimat
Energia elèctrica	0.13343 €/kWh [1]	405 h	31.59 €
Transport	0 €	4 mesos	0 €
Accés a internet	20 €/mes (fora de la universitat)	4 mesos	80 €
Espai en disc (Cloud UPC)	5.75 €/mes	4 mesos	23 €
Total			134.59

Taula 4 Costos generals del projecte

3.1.5 COSTOS IMPREVISTOS

Els costos imprevistos poden provenir de dos fonts principals: Modificacions en els requisits del client que comportin aparicions de noves funcionalitats importants que augmentin la càrrega de treball (amb un risc d'ocórrer d'un 15%) o bé un imprevist greu en les instal·lacions tècniques en les quals es desenvolupa el sistema, ordinador personal o servidors de hosting. (Un risc estimat en total també d'un 15%).

D'altra banda també existeixen riscos a considerar en el fet que la eina DHIS2 és externa a l'equip de desenvolupament, open-source i amb molta activitat contínua. Qualsevol error o canvi mitjanament gran entre versions de la eina ens podria fer alterar els plaços d'entrega i per tant també incrementar els costos.

Finalment s'han estimat uns costos imprevistos no superiors a **200 €** tenint en compte les baixes probabilitats dels riscos i la extensió relativament petita en costos del projecte.

3.1.6 COSTOS TOTALS DEL PROJECTE

Finalment aquest és el resum dels costos del projecte, en tots els casos anteriors i per tant també en el càlcul del cost total, l'IVA i altres taxes ja han estat incloses. Cal esmentar que no s'inclou cap marge de beneficis sobre el cost total del projecte, i les possibles pujades de preu en les diferents tarifes usades no es preveu que tinguin un efecte significatiu.

CONCEPTE	Cost
Costos directes	1948.5 €
Costos indirectes (software, hardware i generals)	296.99 €
Contingència (15%)	297.67 €
Costos imprevistos	200 €
TOTAL	2743.16 €

Taula 5 Informe de costos totals del projecte

3.2 CONTROL DE GESTIÓ

La control de gestió es durà a terme a cada iteració de forma que s'avaluarà la feina realitzada amb la càrrega de treball d'una iteració. D'aquesta forma la càrrega assignada a les iteracions es modificarà en funció del rendiment en el treball.

Al final de cada iteració es realitzarà un control constant d'anàlisis dels costos finals reals, tenint en consideració les hores totals dedicades finalment i comparant aquest valor amb els valors estimats en les seccions anteriors. L'estudi també realitzarà un petit anàlisis per esbrinar la procedència de les variacions amb la planificació, ja siguin a l'alça o a la baixa i s'intentarà corregir per la conseqüent aplicació en les següents iteracions. Si finalment el pressupost total es queda curt, la partida de contingència s'haurà d'aplicar per equilibrar el pressupost. El càlcul de desviacions es realitzarà de la següent manera:

- Desviació de recursos humans en preu = *(Cost plan – cost real) * hores treballades reals*
- Desviació d'una tasca en preu = *(Cost plan – cost real) * hores treballades reals*
- Desviació d'un recurs en preu = *(Cost plan- cost real) * consum real*
- Desviació total en la realització de tasques = *cost total plan tasca - cost total real tasca*
- Desviació total en recursos = *cost total plan recursos – cost total real recursos*
- Desviació total costos fixes = *cost total costos fixes pressupost – cost total fixe real*

Finalment s'utilitzarà la següent taula per calcular les desviacions per cada iteració definides:

Fase del projecte	Tasca	Hores Planificades	Hores reals	Desviació en hores	Cost per hora	Desviació en €
Anàlisi previ i introducció	Inception	105	105	0	4.9 €	-
	Total	105	105	0	4.9 €	-
Disseny i extracció de dades	Sprint 2	55	75	20	4.9 €	98 €
	Sprint 3	55	110	55	4.9 €	269.5 €
	Total	105	180	75	4.9 €	367.5 €
Disseny i implementació del sistema d'injecció de dades al WISCC	Sprint 4	55	65	10	4.9 €	49 €
	Sprint 5	55	65	10	4.9 €	49 €
	Total	105	125	20	4.9 €	98
Avaluació final	Sprint 6	55	65	10	4.9 €	49
	Total	55	65	10	4.9 €	49
Projecte sencer		405	510	105	4.9 €	514.5 €

Taula 6 Càlcul de desviacions per tasca del projecte

Per tant als 2743.16 € de pressupost inicial del projecte hem d'afegir el cost de les desviacions ocasionades durant el desenvolupament d'aquest de 514.5 €. El projecte finalment ha tingut un cost de 3257.66 €.

4. DISSENY DEL SISTEMA

A continuació es presenta el disseny conceptual del sistema, s'ha realitzat de forma dinàmica i mantenint reunions per definir els requisits i validacions dels clients, finalment s'ha definit la següent arquitectura del sistema a desenvolupar. Com ja hem introduït en punts anteriors, el projecte clarament es desglossa en dos grans objectius o parts importants a assolir: El sistema d'extracció de la informació de les instàncies DHIS2 i el sistema d'ingestió d'aquestes dades en el repositori central de dades WISCC.

4.1. SISTEMA D'EXTRACCIÓ

El sistema d'extracció serà el procés software que s'encarregarà d'obtenir totes les noves dades introduïdes en el mòdul d'entrada de dades i enviar-les al sistema central. El seu funcionament bàsic es veu representat a la següent figura:

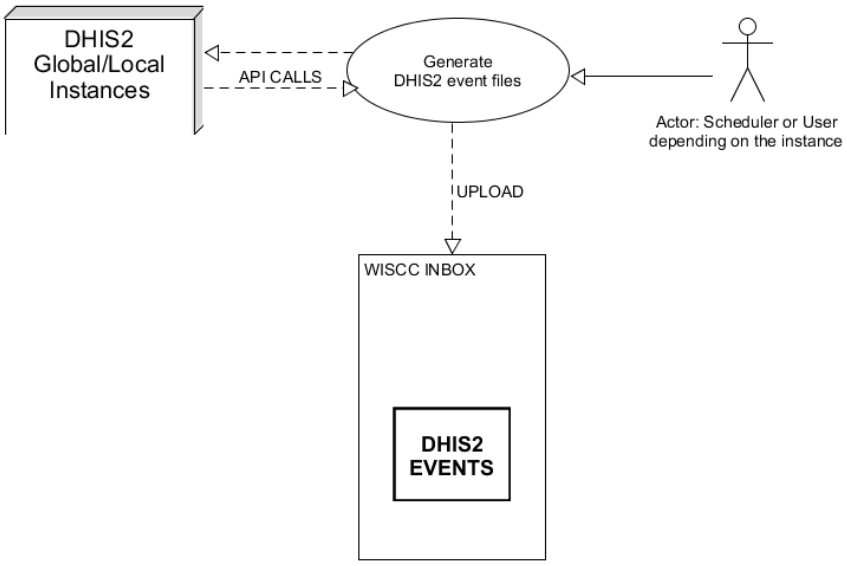

Figura 4 Disseny sistema d'extracció

Com es pot veure a la figura 4, la principal diferència en el tractament de les instàncies Locals i Global de DHIS2 resideix en l'acció que desencadenarà el procés d'extracció, per una banda s'ha definit per les instàncies locals que l'acció estarà lligada a una transacció amb l'usuari i per les instàncies globals aquesta acció es farà de forma automàtica i planificada. El disseny i desenvolupament del mòdul d'entrada o Inbox del sistema WISCC s'ha desenvolupat de forma separada a aquest projecte i simplement s'han coordinat els mètodes i configuracions necessàries per realitzar la transferència de dades.

La generació dels fitxers a transferir segueix el següent disseny o flux d'acció:

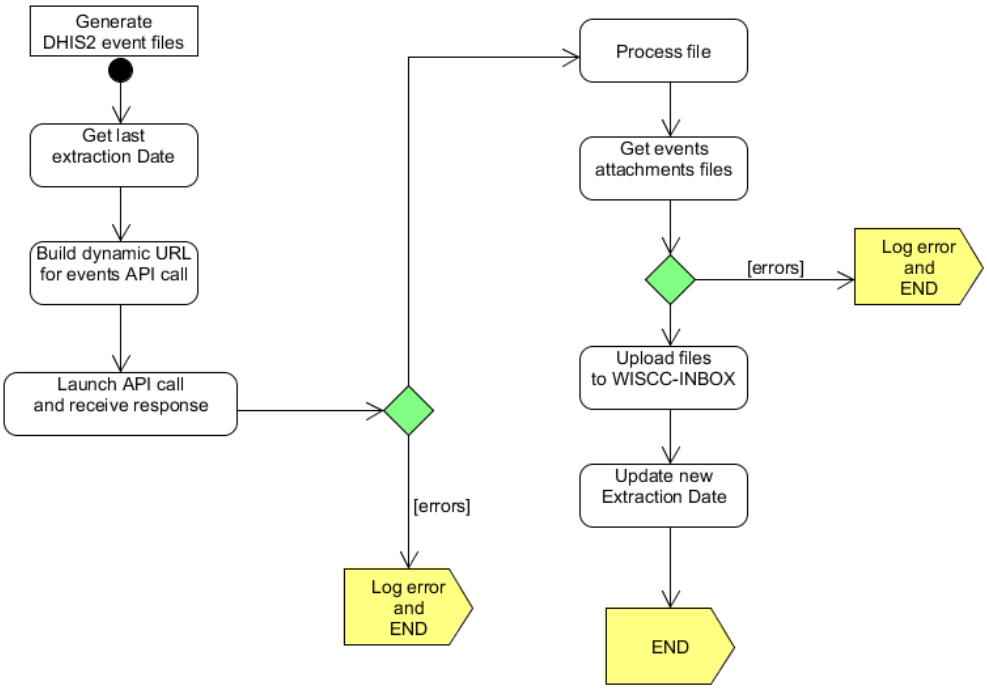

Figura 5 Flux del procés d'extracció

Per entendre correctament el sistema d'extracció es va elaborar també un graf d'estats que indica de forma detallada els estats que les dades introduïdes en el sistema, en un primer moment dins l'eina DHIS2, tindràn. Aquest detall en els estats de la informació introduïda en el sistema és important, no tan sols des d'un punt de vista tecnològic per controlar i executar els processos necessaris en els canvis d'estats pertinents, sinó també a l'hora de mantenir un control de la privacitat i domini de les dades existents. En el cas de la instància del mòdul d'entrada de dades global no hi haurà aparentment problemes referents aquest tema, però els països interessats en mantenir la seva pròpia instancia localment tenen requisits de privacitat i pertinença de les dades introduïdes fins que ells autoritzin el traspàs.

Aquest graf es mostra a la següent figura:

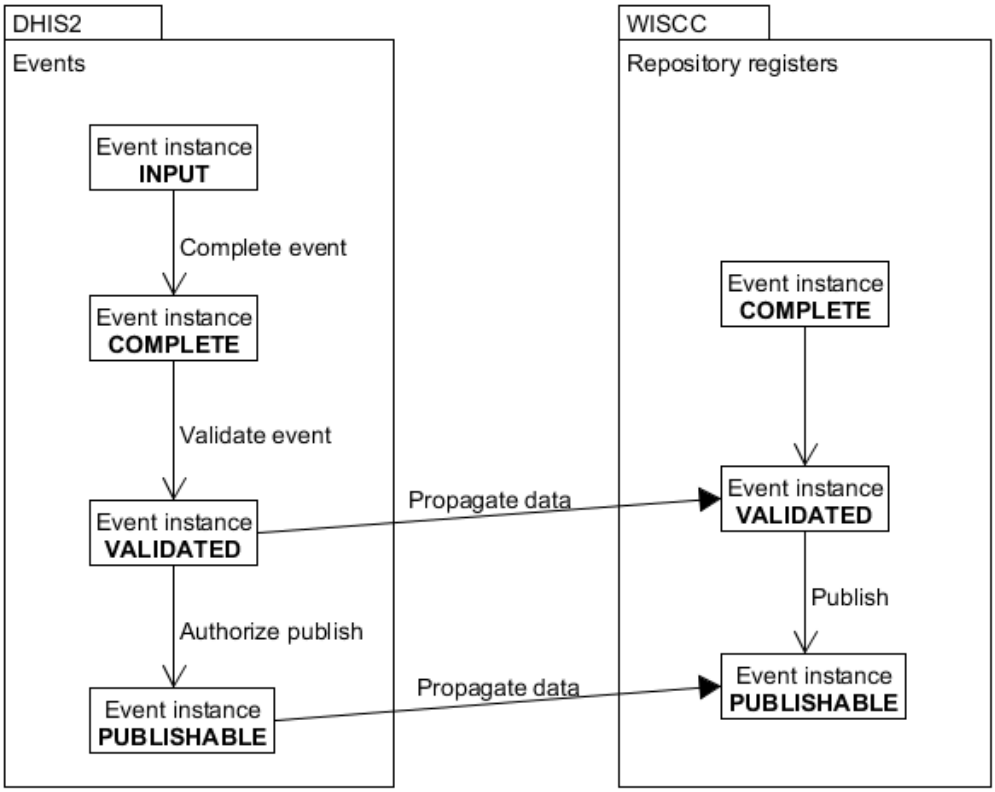

Figura 6 Diagrama d'estats de les dades en el sistema WISCC

Com podem veure en el diagrama d'estats, els individus que tenen estat són per una banda els events afegits a les instancies de DHIS2 i per l'altra els seus correlatius registres en el repositori WISCC. Inicialment les dades estan en un estat d'edició *INPUT* en el qual no es mouen de la instància DHIS2. Aquestes dades que s'agruparan a nivell de formulari o event es poden completar via el responsable d'edició dels formularis i llavors passarien a l'estat *COMPLETED.* En aquest estat les dades són disponibles pels validadors i responsables de les dades de les dades, el rol de validador és aquell que s'encarrega de assegurar una correctesa científica de les dades afegides. El validador pot indicar que ha revisat les dades marcant-les en l'estat de VALIDATED. Quan les dades s'han validat ja es poden propagar cap al repositori central WISCC, en el cas de la instància global de DHIS2 aquesta acció es durà a terme de forma automàtica i en les instàncies locals de cada país es requerirà que un usuari amb rol d'administrador de la instància accioni la transacció que envia tota la informació validada.

Per motius de gestió de la Organització Mundial de la Salud, existeix un quart estat que anomenem PUBLISHABLE, que fa referència a si aquestes dades són de caire confidencial o bé poden ser públiques si la OMS ho creu oportú.

4.2. SISTEMA D'INGESTIÓ DE LES DADES

El sistema d'ingestió de les dades és el procés software encarregat de processar els fitxers que es van incorporant a la Inbox del sistema WISCC i enregistrar-los a la primera capa del Data Lake. El sistema WISCC tindrà un registre configurat amb els diferents fitxers que li són enviats. El sistema d'ingestió consumirà aquest registre per anar tractant un per un tots els fitxers a inserir en el repositori.

La ingestió de les dades es produirà de forma procedimental i intentant aconseguir el màxim d'independència respecte a l'estructura de la font d'entrada. El sistema obtindrà l'estructura dels fitxers d'entrada mitjançant una definició ontològica de les dades que estarà guardada en el sistema mitjançant un graf RDF. La consumició d'aquest graf permetrà al sistema conèixer tots els procediments, dades i execucions que ha de dur a terme per inserir correctament els fitxers al repositori. D'aquesta manera, qualsevol canvi en l'estructura tant dels fitxers com de la capa primària del repositori es veuran reproduïts a l'ontologia i el sistema no es veurà afectat en cap moment.

L'esquema general del sistema d'ingestió de les dades es representa a la següent figura 7:

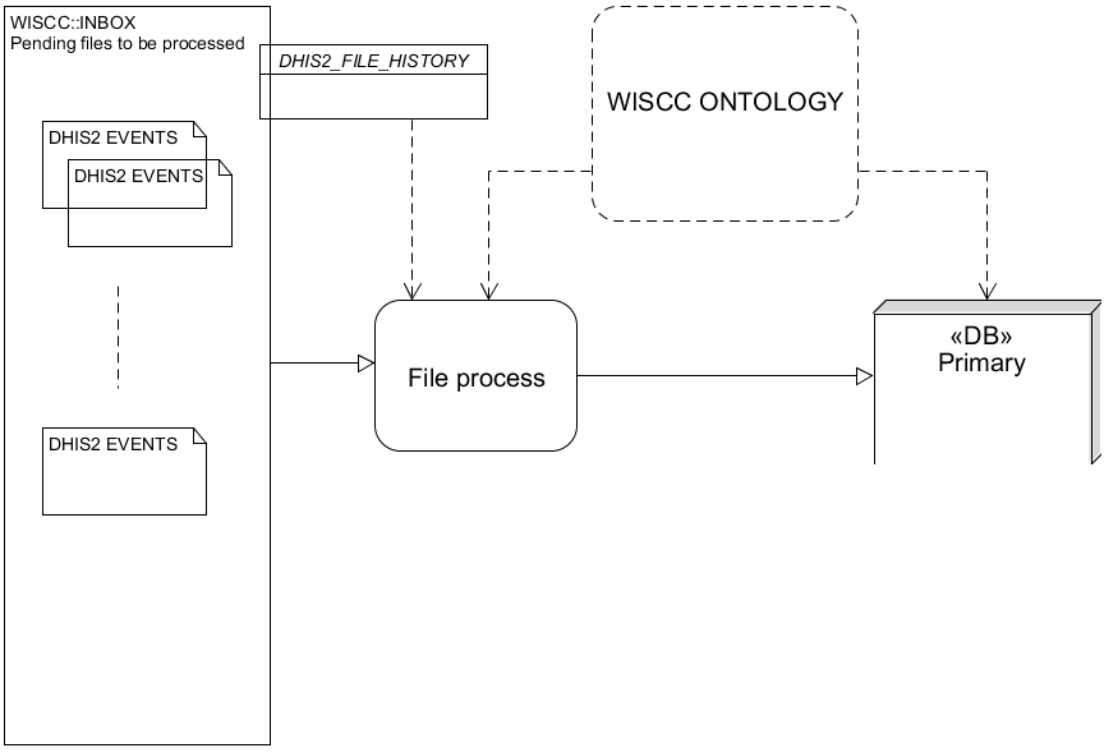

Figura 7 Esquema general del sistema d'ingestió de les dades

Com podem veure, els fitxers d'entrada procedents tant de la instància global com de les locals estaran emmagatzemats dins de la Inbox del sistema WISCC i contindran una anotació específica a les Bases de Dades de control EVENT_FILE_HISTORY per tal d'identificar el fitxer, la seva data d'arribada i el seu estat actual (pendent o no d'ingestió).

El procés d'ingestió consumirà la base de dades per saber quins fitxers ha de processar i realitzarà la ingestió de cada un d'ells. Per realitzar la ingestió s'alimentarà d'un esquema ontològic que definirà l'estructura de les dades procedents de l'eina DHIS2. Aquesta ontologia formarà part del gestor de metadades del sistema WISCC i ha estat desenvolupat interiorment en el projecte WISCC per tal de poder abstraure del sistema d'ingestió tota mena de dependència amb l'estructura de les dades a inserir.

4.2.1. ONTOLOGIA DEL CANAL AMB DHIS2

Dins del projecte WISCC s'ha definit l'objectiu principal de dissenyar una ontologia de coneixement del sistema que enllaci i dirigeixi els coneixements des de les capes més pròximes als usuaris finals fins les capes del coneixement més pures i tècniques de les bases de dades. Una part d'aquesta ontologia per tant ha de donar suport a les dades específiques per poder tractar els fitxers provinents tant del sistema DHIS2 com d'altres sistemes externs i poder alimentar el repositori sense cap dependència de l'estructura de les dades dels sistemes exteriors. Per realitzar això s'ha definit un mapeig entre els conceptes específics de l'eina DHIS2 i els conceptes propis de l'esquema del repositori central.

El desenvolupament tant teòric com pràctic d'aquesta ontologia no forma part de l'abast d'aquest projecte, és per això que no es realitza un estudi detallat i justificat del seu disseny. Ha estat definida mitjançant un graf RDF en el qual es defineixen totes les entitats existents i les seves relacions (entre elles) i propietats. Inclusivament en aquestes definicions també s'inclouen regles de validació i passos extres dins del procés d'ingestió intrínsecs de cada una de les entitats.

L'objectiu final és aconseguir que mitjançant aquest graf el sistema d'ingestió de les dades pugui processar un fitxer pendent d'ingestió sense necessitat de conèixer la seva estructura ni les seves validacions. Com ja hem introduït les principals components de la ontologia que participen en el processament dels fitxers provinents de les instàncies de DHIS2 són:

- La definició del canal amb DHIS2, és a dir, totes les definicions per poder entendre l'estructura de dades de la font d'informació.
- La definició de l'esquema intern de la primera capa del repositori del sistema WISCC, el target de la ingestió de les dades.
- I finalment el mapeig, és a dir, l'enllaç entre els conceptes d'una estructura amb l'altra per tal de poder obtenir insercions al repositori a partir dels fitxers processats. A partir de les dades dels fitxers poder inserir dades en el repositori central. És el cas per exemple dels programes de DHIS2 enllacen directament amb els Data sets de l'esquema d'integració i els elements dels programes són directament els atributs dels Data sets.

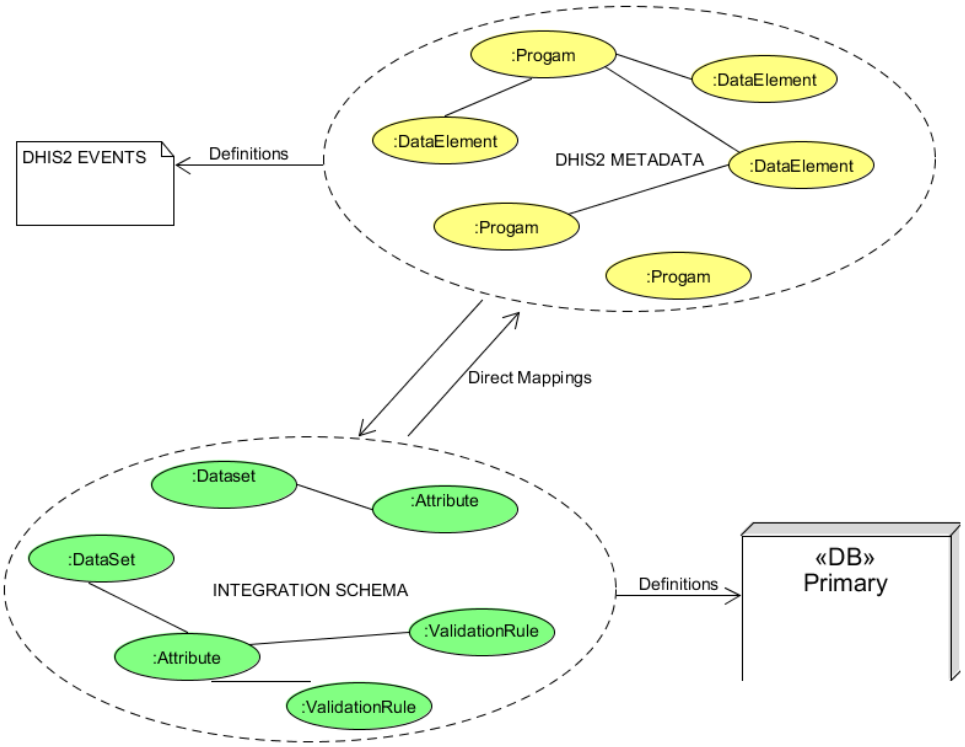

Figura 8 Funcionament de la gestió del canal DHIS2 mitjançant les ontologies

4.2.2. DISSENY DE LA PRIMERA CAPA DEL REPOSITORY CENTRAL

L'etapa final d'inserció en la capa primària del repositori WISCC és una etapa oberta en la qual es manté una constant adaptació als requisits de les ontologies en desenvolupament. Tot i això s'ha realitzat un disseny previ conceptual pel que fa referència a l'estructura interna de les dades que s'emmagatzemaran.

La tecnologia elegida per implementar aquesta primera capa s'ha decidit per part de tots els components del projecte que sigui la base de dades Document-Store Mongo DB, en concret una de les seves últimes distribucions a partir de la Mongo 3.0 ja que incorporà un nou motor gestor de base de dades WiredTiger que proporciona índexs en arbres i realitza consultes per rang.

En concret Mongo DB és un bon candidat ja que permet un esquema de dades flexible i adaptable als canvis que el sistema WISCC preveu i treballa nativament amb fitxers Bson, la versió binària dels fitxers JSON, els quals són molt presents en tot el desenvolupament del sistema. Finalment es preveu també que les seves aptituds tècniques siguin beneficioses pel desenvolupament actual i futur del sistema: consultes en rang i aleatòries, nombroses APIs en diferents llenguatges de programació, APIs REST via HTTP implementades i gran comunitat de desenvolupadors.

MongoDB és un motor gestor de bases de dades Document-Store, les bases de dades orientades a documents estan constituïdes per un conjunt de programes o col·leccions que emmagatzemen dades de documents amb una certa estructura. Un document és una entitat que encapsula informació seguint un cert format estandarditzat, ja siguin XML, YAML, JSON O BSON. Podríem dir que els documents són similars al concepte de registre o fila en les bases de dades relacionals. La gran diferència és que no tenen un esquema fixa que els governi ni en els atributs, claus o altres seccions. Els atributs del document és el concepte que es podria emparellar millor amb el concepte de columna o atribut a les bases de dades relacionals.

Els formularis que es presenten en el mòdul d'entrada de dades mitjançant DHIS2 es presenten agrupades en programes. Cada programa té un formulari específic i pot rebre tant registres com formularis es completin. Pel que fa als programes, aquests s'agrupen en 3 grans grups (Packages): Systemic, Healthcare i Transmission Interrupt.

Així doncs basarem la primera decisió d'organització d'aquesta capa amb aquests conceptes. Els 3 packages representaran 3 directoris diferents dins d'una mateixa jerarquia de fitxers. Per cada directori, les dades es dividiran en col·leccions, una per cada programa, i a dins de cada col·lecció un registre representarà la informació que s'ha introduït completant un formulari. Cada document tindrà una clau única que l'identificarà dins de la col·lecció.

Altres aspectes més tècnics de l'organització dels registres a la capa primària del repositori MongoDB són primordial per les següents raons:

- S'ha de mantenir una gestió de les versions dels registres de la base de dades
- Obtenir la última versió del registre ha de ser còmode i eficient.
- L'actualització i per tant la creació d'una nova versió ha de ser consistent i eficient, tant en complexitat com en memòria ocupada.
- Si el sistema fallés en qualsevol moment, necessitem tenir sempre un estat consistent del repositori.

Seguint aquestes directrius es van dissenyar 3 possibles candidats com a esquemes conceptuals dels registres de la primera capa del repositori:

a. Guardar les dades del formulari complertes cada vegada que es rep una nova versió del registre. Afegint un atribut més temporal com a Timestamp que ens indicarà la seva data de creació en el repositori.
b. Guardar totes les versions del registre dins d'un mateix document, diferenciant la versió actual de les versions prèvies.
c. Guardar per cada registre tants documents com modificacions entrin en el repositori però només guardar els deltes d'aquestes modificacions, és a dir, la informació nova o actualitzada.

L'opció A és possiblement la més consistent ja que per cada versió tens un document complet i suficient. De totes formes, és l'opció que ocupa més memòria i obtenir la versió actual significaria llegir tots els registres per trobar el timestamp més actual. L'opció B, en canvi, conté tota la informació en el mateix document i l'obtenció de la versió actual es pot aconseguir amb un únic accés, tot i això l'actualització d'una nova versió seria més costosa degut a que s'hauria de modificar el document, col·locar la versió actual com a prèvia i inserir la nova versió com a actual. MongoDB no disposa d'un bon tractament de les actualitzacions dels registres degut al sistema amb el que guarda els fitxers i provoca un malbaratament

de memòria perjudicial pel sistema. Finalment l'opció C no s'adaptaria a la flexibilitat que el projecte contempla en el seus objectius principals i generar la versió actual o versions en un estat previ serien costós.

A partir d'aquestes justificacions es va decidir optar per la opció A, una opció que accepta la redundància de la informació i possibles valors nuls en el cas que els events no tinguin totes les dades afegides quan ja s'insereixin en el sistema però que prioritza l'estabilitat del sistema i la capacitat d'obtenir la versió actual dels documents o les versions en un estat previ determinat de forma fàcil i eficaç.

Així doncs, finalment dins de cada programa o col·lecció de la base de dades cada registre representarà l'estat d'un formulari en un cert temps. Per tant la clau primària d'un dels registres serà composta i es compondrà de l'identificador del event (únic dins de la col·lecció) i el seu timestamp. D'aquesta manera un event tindrà tants registres a la base de dades com vegades s'hagi importat informació sobre aquest en el sistema, per obtenir la versió actual simplement es requerirà consultar tots els events amb l'identificador determinat i elegir el que tingui l'atribut temporal més proper a la data actual.

El desenvolupament d'aquesta capa primària del repositori serà primordial pel sistema WISCC per posteriorment afegir eines analítiques i dimensionals per aconseguir coneixement a partir de la informació emmagatzemada en el Data Lake.

5. DESENVOLUPAMENT DEL SISTEMA

El sistema s'ha desenvolupat a partir del disseny establert intentant prendre les decisions que afavorissin al màxim les directrius de dissenys i objectius més rellevants del projecte WISCC. Tot el desenvolupament s'ha realitzat amb Java ja que és un llenguatge tipat amb el qual els desenvolupadors tenen força experiència i complia els requisits necessaris definits en el disseny.

5.1. DESENVOLUPAMENT DEL SISTEMA D'EXTRACCIÓ

5.1.1. VISIÓ GENERAL

Una de les principals directrius de desenvolupament que s'han aplicat en el projecte és la de la reutilització de codi i abstracció total dels processos i dades de configuració. És per això que s'ha intentat extreure al màxim la lògica del sistema d'extracció per tal de no haver de generar dos processos àmpliament independents per tractar tant les instàncies locals com la instància global.

Amb aquest propòsit s'ha implementat el sistema DHIS2-CONNECTOR, un sistema Java que realitza a partir de una sèrie de fitxers de configuració l'extracció de dades de la instància en concret i transfereix els fitxers generats a la INBOX del sistema WISCC. Aquest procés Java serà l'eix essencial per generar l'extracció tant per les instàncies locals com la global. DHIS2-CONNECTOR encapsula les funcionalitats necessàries per realitzar una extracció i s'utilitza en els dos mètodes de desplegament per cada una de les instàncies del mòdul d'entrada de dades.

5.1.2. DHIS2 CONNECTOR

Com ja hem introduït en el punt anterior, DHIS2-CONNECTOR és el motor principal del sistema d'extracció . El seu flux d'operacions és exactament el presentat a la **figura de flux del sistema d'extracció. (figura 5)**

5.1.2.1. FITXERS DE CONFIGURACIÓ

Conté tres principals fonts de configuració:

- Per una banda conté tota la informació necessària pel control de l'última data d'extracció. Aquest punt és important per mantenir una extracció incremental de les dades introduïdes a l'eina DHIS2.
- Informació de configuració necessària per generar dinàmicament la URL que ens permetrà recollir totes les dades introduïdes mitjançant operacions amb la API de DHIS2.
- Per últim també conté totes les configuracions necessàries per connectar amb el sistema WISCC i realitza la transferència.

5.1.2.2. UTILITZACIÓ DHIS2 WEB API

L'objectiu principal és extreure tots els formularis introduïts a l'eina DHIS2 i que encara no han estat registrats en el repositori central del sistema WISCC. És a dir, o bé perquè han estat creats posteriorment a la última extracció o bé modificats. A continuació expliquem els paràmetres utilitzats per construir dinàmicament la URL necessària per obenir les dades a extreure mitjançant la API web de DHIS2.

Tant a la instància global com a les instancies locals els formularis de DHIS2 tenen un atribut temporal que ens indica la seva última data de modificació: **lastUpdatedDate.** Per completar les condicions que han de complir els formularis que es volen registrar en el sistema central també han de ser aquells que ja estiguin factibles de ser extrets, és a dir, que l'usuari ja hagi indicat que totes les dades ja han estat introduïdes. Veurem en els següents punts que el valor que es filtra difereix depenent de la instància que estiguem tractant.

Per últim, l'eina DHIS2 també obliga a indicar sobre quin programa es vol extreure les dades. Degut a aquest paràmetre el procés primer de tot obté tots els programes que estan configurats a l'eina i realitza tantes crides com programes amb els paràmetres presentats anteriorment.

A la crida final per obtenir els events també s'han afegit paràmetres de configuració com **skipPaging** per evitar que la resposta vingui separada amb parts. Finalment el sistema utilitza la seguent sintaxi per realitzar les crides necessàries per l'extracció, en aquest cas busquem dades que s'hagin indicat en estat *COMPLETED*:

(SERVIDOR)/api/events.json?program=<program_id>&lastUpdated=2016-01-01&status=COMPLETED&skipPaging=true

On SERVIDOR fa referència a la localització dels de la instància dhis2, host i port o nom DNS.

5.1.2.3. ESTRUCTURA DE LA RESPOSTA

L'adreça URI presentada en el punt anterior proporciona la següent estructura de resposta:

```
{"events": [{   //Vector amb els events que han complert les condicions d'extracció
 "programStage": "wv2EymKDQCB",  //Identificador intern DHIS2
 "dataValues": [],  //Valors dels formulary que s'han introduït/modificat.
 "storedBy": "admin",  //Usuari que han enregistrat l'event
 "notes": [],
 "created": "2016-02-18T09:54:11.748+0000",  //Data de creació
 "orgUnit": "uZZhXR5xxmV",  //Unitat organitzativa associada al formulari
 "dueDate": "2016-02-18T10:05:28.576+0000",
 "program": "BjNDPeUcTq8",//Programa associate al formulari extret
```

```
 "completedDate": "2016-02-16T23:00:00.000+0000", //Data en que es va completar
 "lastUpdated": "2016-02-18T10:05:28.577+0000", //Data de modificació
 "href": "http://who-dev.essi.upc.edu:8080/api/events/dNXr4iCXEsf",
 "event": "dNXr4iCXEsf", //identifador intern de l'event
 "completedBy": "admin", //Usuari que va completer l'event
 "status": "COMPLETED", //Status
 "eventDate": "2016-02-16T23:00:00.000+0000",
 "orgUnitName": "WHO-HQ" //Nom de la unitat organitzativa associada al formulari
}]}
```

5.1.2.4. CRIDES ADDICIONALS

Per evitar la presència d'identificadors interns de la instància de DHIS2 en els fitxers extrets, el sistema tradueix aquests identificadors a identificadors oficials del projecte WISCC conforme a la definició de les metadades del sistema. Aquest procés es realitza mitjanánt dues crides addicionals:

- Per una banda es consulten tots els programes que la instància de DHIS2 en qüestió conté i es realitza una traducció entre el identificador de programa (identificador intern) amb el seu atribut ShortName (identificador global del sistema WISCC) que haurà estat inicialment configurat.
- De forma anàloga es realitza la traducció entre els identificadors interns dels Data Elements de DHIS2 amb el seu atribut Name, que també haurà estat afegit seguint amb les directrius de nomenclatura del sistema global WISCC.

5.1.2.5. TRANSFERÈNCIA A WISCC-INBOX

Finalment el sistema transfereix aquests fitxers a l'entrada Inbox del projecte WISCC on seran anotats per processar. La inbox del sistema WISCC s'ha configurat com un Servei Web en el qual es realitza un post en el següent recurs:

(SERVIDOR)/JAXRS-Inbox/upload/upload

El mètode post té 4 paràmetres:

- File: Fitxer en format JSON amb totes les noves dades o modificades presents en l'extracció
- zipFile: Fitxer .tar.gz amb els documents adjunts dels formularis extrets
- creationModality: Identificador del tipus d'instància DHIS2 {local | global}
- sc: Codi d'identificació per seguretat

5.1.3. EXTRACCIÓ A LA INSTÀNCIA GLOBAL: CRON JOBS

Una vegada explicat el funcionament DHIS2-CONNECTOR, el seu acoblament a les diferents instàncies depèn majoritàriament de com s'executa el procés. En aquest cas, com hem explicat en els punts introductoris del document, la instància global del sistema és aquella que residirà en els dominis de l'Organització Mundial de la Salut, la qual per tant té total propietat de les dades una vegada introduïdes a la eina. És per això que el procés d'extracció no requereix de cap mena de consentiment per part de l'usuari. Una vegada el formulari s'ha indicat com a *COMPLETED*, aquest ja és factible de ser extret cap als sistemes del projecte WISCC. De totes formes s'ha de mantenir un registre de quines dades complertes ja estan validades o no per tal de poder determinar si poden passar a estat *PUBLISHABLE*. És per això que en la instància global l'execució dels procediments del DHIS2-CONNECTOR es realitzarà de forma automàtica i planificada mitjançant CRON JOBS. CRON és un servei de planificació de tasques basat en temps molt utilitzat en sistemes operatius Unix, aquest permet executar comandes de forma automàtica en un temps específic.

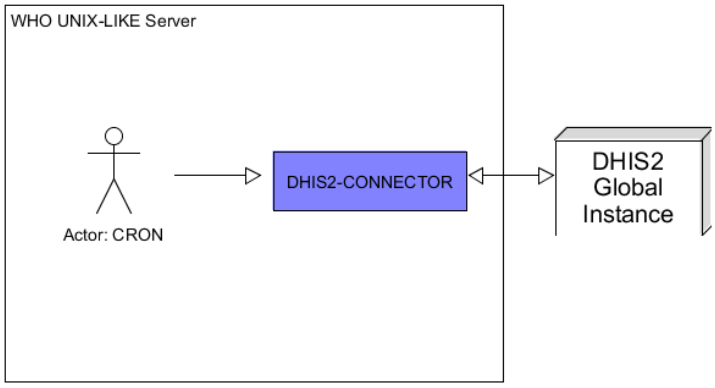

Figura 9 Execució DHIS2-CONNECTOR en la instància global

5.1.4. EXTRACCIÓ EN INSTÀNCIES LOCALS: RESTFUL WEB SERVICES

En les instàncies locals ja hem explicat que l'eina DHIS2 estarà gestionada pels mateixos administradors dels propis i països i per tant les dades que s'introduiran seran propietat del país. És per això que la seva extracció requereix d'un tractament superficialment diferent que consisteix en requerir d'una acció de l'usuari per iniciar el procés d'enviament de les dades validades als repositoris centrals del sistema WISCC.

És per això que en aquest cas s'ha implementat un servei web REST que permet per una banda iniciar l'execució del procés DHIS2-CONNECTOR i per altra consultar en format CSV quines dades s'estan a punt de extreure.

Aquest servei web serà consumit mitjançant una aplicació desenvolupada pels desenvolupadors del sistema WISCC i permetrà a l'usuari autoritzar l'extracció i l'enviament de les dades.

Com es pot veure a la figura 7 el servei web tindrà 4 endpoints o recursos principals:

- */events/summaryCSV* : Representarà el resum en format CSV dels events que s'exportaran a la pròxima extracció. L'usuari es podrà baixar aquest fitxer.
- */extraction* : Realitza una extracció i retorna un codi indicant el seu resultat (Error/Extracció buida/Extracció correcte)
- */events/data-summary* : L'usuari es podrà baixar el fitxer .json que s'ha enviat a la última extracció.
- */events/data-attachments :* L'usuari es podrà baixar el fitxer comprimit amb tots els fitxers que s'han adjuntat als events extrets en la última extracció.

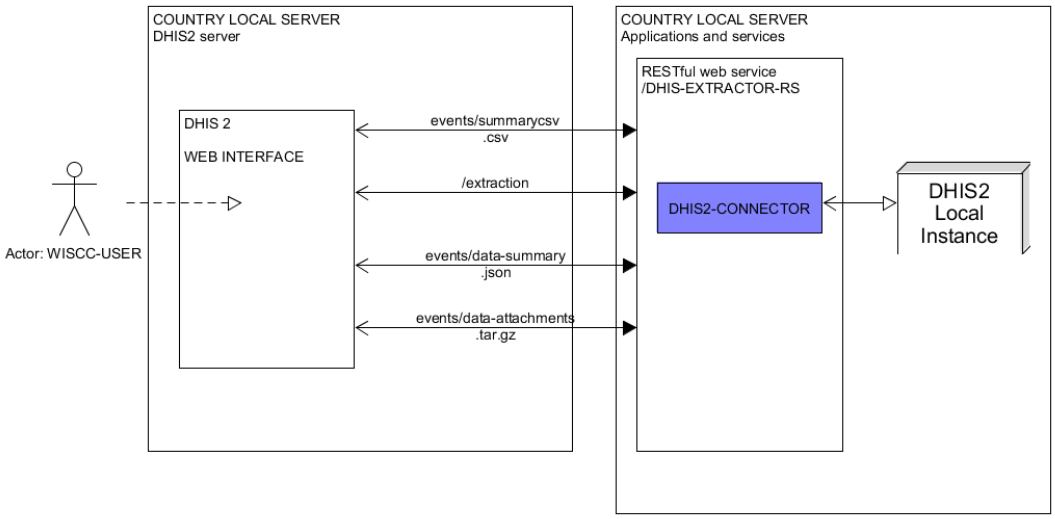

Figura 10 Sistema d'extracció per instàncies locals

5.2. DESENVOLUPAMENT DEL SISTEMA D'INGESTIÓ

El disseny del sistema d'ingestió de dades al sistema WISCC s'ha especificat en l'apartat de disseny amb detall. En l'etapa de desenvolupament del sistema d'ingestió simplement ha significat el desenvolupament d'un programa Java que consumís els fitxers afegits al sistema WISCC provinents de la extracció de la informació de les instàncies de l'eina DHIS2 i el seu tractament i final inserció mitjançant les definicions i estructures definides per l'Ontologia del projecte WISCC.

Aquest desenvolupament no es considera finalitzat a l'entrega d'aquest document ja que la seva construcció i també el disseny de la Ontologia de les metadades del sistema es consideren uns processos àgils i canviables que depenen dels requisits dels clients. Una vegada la gestió de les metadades s'hagi realitzat completament el sistema d'ingestió estarà totalment desenvolupat i serà capaç de tractar tot tipus de fitxers i dades que provinguin del canal d'entrada DHIS2 i aplicarà els tractaments indicats per finalment inserir nous registres al repositori.

De totes formes s'ha dissenyat un algorisme com a estructura principal a seguir i punt de referència per anar completant el procés final d'ingestió.

5.2.1. ESTRUCTURA DE L'ALGORISME D'INGESTIÓ

Com ja hem explicat, el procés d'ingestió comença amb els fitxers JSON de diferents extraccions emmagatzemats al mòdul Inbox del sistema WISCC. L'algorisme d'ingestió primer de tot s'encarregarà de consultar a la taula de referència d'aquest mòdul Inbox quins són els fitxers que encara no s'han processat. Un cop obtinguts els fitxers, aquest serà el procediment a seguir per cada un d'ells:

- Obtenir tots els events que formen part del fitxer, és a dir, tots els formularis que en aquella extracció el sistema va decidir extreure. A continuació per cada event es realitza el següent procés:
    - Primer de tot s'obté quin és el programa [DHIS2] al qual pertany el formulari extret. Aquest programa enllaça directament mitjançant l'Ontologia del sistema WISCC amb una col·lecció del repositori.
    - Mitjançant les dades del programa definides en l'esquema d'integració, el procés ja pot crear l'identificador únic que s'inserirà com a clau primària en el repositori MongoDB. Aquest procés també reconeixerà si el nou formulari és un fitxer nou en el repositori o és una actualització d'un ja existent.
    - Tot seguit s'obtenen les dades (files del formulari) que s'indiquen com a afegides o modificades. Per cada una de les dades es realitza una validació mitjançant dues parts essencial: Primerament es valida si aquesta dada està associada al programa per mitja de l'ontologia del canal DHIS2 i en segon lloc es realitzen totes les validacions semàntiques i sintàctiques de la dada en qüestió, definides també mitjançant l'ontologia del canal. Si alguna dada no es validada correctament s'anota l'error però no es cancel·la el procés d'ingestió.
    - Una vegada validades totes les dades del formulari també es validen les dades intrínseques del formulari com serien el seu estat, usuari que l'ha actualitzat, dates de creació i unitat organitzativa (País o regió) a la que pertany. Qualsevol validació errònia de les dades també s'anotarà.

- Si la dada a validar és un fitxer, ja hem introduït anteriorment que els formularis poden tenir fitxers adjunts com a informació extra, el procés també validarà el fitxer adjunt i l'inclourà en el sistema, afegint al registre del repositori una clau forana en forma de ruta relativa de fitxers.

- Una vegada processat l'event es consulta el registre d'errors. Si hi ha algun error, el procés es cancel·la l'extracció de tot el fitxer i s'activa el tractament d'errors en fitxers d'ingestió que s'explicarà en el següent apartat.
- Si no hi ha cap error es procedeix a inserir el nou o la nova versió del event en el repositori del projecte.

5.2.2. TRACTAMENT DELS ERRORS EN EL PROCÉS

A l'algorisme d'ingestió hem indicat que s'anotaran tots els errors produïts les validacions de les dades. Un cop finalitzat el procés d'ingestió, qualsevol d'aquests errors provoca la cancel·lació de la ingestió del fitxer sencer. Tot i això el procés no s'acaba aquí, el procés d'ingestió anota l'error i deixa el fitxer en un estat de pendent de revisar en el qual es requerirà d'una transacció d'usuari per analitzar l'error, solucionar-lo i afegir el fitxer en l'estat de pendent de processar com si acabés de ser extret dins del mòdul Inbox del sistema.

6. INFORME DE SOSTENIBILITAT

A continuació es detalla l'informe de sostenibilitat d'aquest projecte. L'anàlisi s'ha realitzat des del punt de vist de tres dimensions: Econòmica, Social i Ambiental. Aquest anàlisi es veu resumit en la següent matriu de sostenibilitat del projecte:

	PPP	Vida útil	Riscos
Ambiental	8	16	0
Econòmic	9	16	0
Social	10	20	0
Rang de sostenibilitat	27/30	52/60	0 (entre -60 i 0)
		79/90	

Taula 7 Matriu de Sostenibilitat del projecte

Els raonaments i justificacions de les puntuacions anteriors venen definides en les següents seccions.

6.1. DIMENSIÓ ECONÒMICA

S'ha definit clarament una avaluació de costos materials i humans i tant en la planificació com en el seu posterior desenvolupament es mantindran mètodes d'ajustament i gestió del riscos. Els temps dedicats a cada tasca són proporcionals a la seva rellevància dins del projecte i forma part d'un projecte aprovat i subvencionat per l'Organització Mundial de la Salut dins del Programa de control de la malaltia Chagas.

6.2. DIMENSIÓ SOCIAL

La dimensió social és un dels punts més positius del projecte degut al camp d'actuació del sistema central on aquest projecte s'instal·larà, el sistema WISCC. Aquest sistema software d'emmagatzematge i anàlisi de dades parteix com a objectiu principal de construir un sistema capaç de proveir d'informació útil en salut sobre una de les malalties endèmiques més propagades en l'actualitat, el Chagas. Amb aquest projecte per tant, es pretén aconseguir una millora social important. Gràcies aquest sistema es podran gestionar molt millor les distribucions de medicaments pels tractaments de la malaltia, les inspeccions a domicilis en risc o l'aparició de brots en zones inesperades entre d'altres.

Dins del mateix problema però des d'un punt de vista molt diferent, aquet projecte també aconseguirà una transparència addicional en els reports de salut dels països. La OMS tindrà la informació en temps real i no dependrà d'informes o articles realitzats en els països que poden passar per processos de censura i tergiversació.

6.3. DIMENSIÓ AMBIENTAL

La dimensió ambiental conté un del principals problemes que tot Data Center o magatzem de dades contempla. L'emmagatzematge de dades presumeix de ser protagonista de molts estalvis tant en infraestructura, com en espai o energia. Tot i això quan el volum de dades augmenta, el manteniment d'aquests magatzems de dades esdevé costós i perjudicial pel medi ambient.

Pel que fa a l'abast detallat del projecte, tant en la seva realització com la seva vida útil no representa costos rellevants medi-ambientals, ja que simplement serà un mecanisme de transport o flux de les dades des d'un port fins un altre.

7. CONCLUSIONS

7.1 ASSOLIMENT DELS OBJECTIUS

La gran majoria dels objectius presentats en un primer moment en la formulació del projecte s'han assolit i de forma molt satisfactòria. Cal remarcar la importància dels objectius de disseny que el projecte WISCC fa referència en totes les seves accions. Sobretot la importància de mantenir el software dissenyat totalment independent de les estructures de dades i el coneixement del sistema. Aquest és un punt molt important ja que el món dels magatzems de dades cada vegada contempla dades més disperses i de moltes semàntiques diferents impossibles d'englobar totes en un mateix i fixe esquema. Per tant poder tenir dades de molts formats diferents i no haver de gestionar cada format amb un software diferent ens generar un speed up productiu i qualitatiu determinant.

Tant en el desenvolupament del sistema d'extracció de les dades com en el sistema d'ingestió d'aquestes al repositori central s'ha intentat mantenir com a principal premissa aquest objectiu. Els processos software no depenen de dades estructurals i en qualsevol cas s'alimenten sempre de fitxers de configuració editables. La responsabilitat s'accentua en el sistema d'ingestió en el qual, com hem explicat, tot el procés s'alimenta de l'Ontologia de coneixement del sistema. Si aquesta Ontologia del sistema WISCC és capaç d'enllaçar-la amb els usuaris i fer-los presents en la creació de les metadades del sistema voldrà dir que aquestes metadades seran les més pròximes i fiables als requisits de l'usuari.

Els objectius del desenvolupament del sistema d'ingestió es van veure truncats respecte a la planificació degut a les nombroses iteracions de validació que ha hagut de passar tant el disseny com la interfície final del sistema d'extracció. Per una banda positiu, perquè vol dir que s'han analitzat curosament els requisits dels usuaris i s'han complert les seves expectatives, però també negatiu per què el sistema d'ingestió clarament encara requereix de refinaments en el seu disseny i implementació.

7.2 CONEIXEMENTS APLICATS I ADQUIRITS

Tal i com es va explicar en l'elecció de les competències d'aquest treball, durant el seu desenvolupament s'han aplicat diversos i nombrosos conceptes relacionats amb l'Enginyeria del Software. Per una banda reafirmant i complementant el coneixements donats com són les premisses de disseny, fluxos d'execució, ús d'APIs per extreure informació, organització i simplicitat de producció.

I per altra banda i potser més important, s'ha tingut la oportunitat d'aprendre i ser protagonista en altres apartats. En primer lloc la participació en reunions amb els clients per poder extreure el màxim de retroalimentació possible. S'ha profunditzat en coneixements dels sistemes ETL per tal de poder mantenir un flux de dades controlat i organitzar els problemes en peces més petites per simplificar. S'han obtingut també nous coneixements pel que fa al disseny d'estructures de bases de dades NoSQL, sobretot a saber analitzar les teves principals característiques com a repositori de dades i a partir d'aquí extreure el motor gestor de base dades que més s'adequa a les necessitats.

7.3 TREBALL FUTUR

El treball futur és clar, el sistema d'ingestió necessita ser més robust i més complet. A mesura que el desenvolupament de l'ontologia també avanci el sistema d'ingestió ha de ser capaç d'explotar tot el coneixement representat en l'ontologia per aplicar tots els processos de validació i ingestió necessaris per convertir les dades provinents del sistema d'extracció en nous registres en el repositori. Per suposat cap de les feines fetes es dona com a acabada ja que el projecte WISCC intenta estar en constant validació i reunió amb els clients per tal de satisfer els seus requisits.

Un altre punt molt interessant i que en un primer moment estava incorporat dins dels objectius d'aquest treball és la generació d'un panell d'informació sobre el flux de dades que intervenen en el sistema d'extracció i ingestió. És a dir, com ja hem explicat en el treball, s'ha mantingut una constant anotació informativa del transcurs de les dades des del sistema d'extracció fins que acaben sent inserides en el sistema. Aquestes dades anotades es poden analitzar per informar al client de les seves característiques (països, mesos de l'any, freqüències, problemes, etc). Les eines de Logging actuals permeten definir una estructura fixe als fitxers d'anotació que ens permet posteriorment mitjançant un sistema software extern analitzar aquests fitxers i dissenyar uns panells d'informació per l'usuari. La creació d'aquest procés software de mineria d'informació a partir dels fitxers d'anotació dels sistemes d'extracció i ingestió seria un dels principals punts en les metes a assolir en un futur proper.

8. REFERÈNCIES

[1] World Health Assembly. [en línia] WHO. [Consultat: 26-Febrer-2016] Disponible a Internet: http://www.who.int/en/

[2] World Health Organization. *Working to overcome the global impact of neglected tropical diseases: first WHO report on neglected tropical diseases*, volume WHO/HTM/NTD/2010.1. World Health Organization, 2010.

[3] World Health Organization. Sustaining the drive to overcome the global impact of neglected tropical diseases. Second WHO report on neglected tropical diseases, volume WHO/HTM/NTD/2013.1. World Health Organization, 2013.

[4] World Health Organisation Staff. WHO — The 17 neglected tropical diseases. 2013.

[5] José Rodrigues Coura and Pedro Albajar Viñas. Chagas disease: a new worldwide challenge. Nature, S6–S7, 2010

[6] Jaume Viñas Navas (2015) *Building a Data Warehouse for the global WHO information and surveillance System to control/eliminate Chagas disease* [en línia] Barcelona: Facultat d'Informàtica de Barcelona. Universitat Politècnica de Catalunya. [Consultat: 28-Febrer-2016] Disponible a Internet: http://upcommons.upc.edu/bitstream/handle/2099.1/20800/92055.pdf

[7] DHIS Organization [en línia] [Consultat: 29-Febrer-2016] Disponible a Internet: https://www.dhis2.org/

[8]Pravin Ganore. *Introduction to the concept of Data Lake and its benefits* [en línia] ESDS.Enabling futurability. [Consultat: 28-Febrer-2016]. Disponible a Internet: http://www.esds.co.in/blog/introduction-to-the-concept-of-data-lake-and-its-benefits/#sthash.h4UZzqoE.dpbs

[9] Brian Stein i Alan Morrison. *The enterprise data lake: Better integration and deeper analytics* [en línia] Techonoloy Forecast. [Consultat: 28-Febrer- 2016] Disponible a Internet: http://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf

[10] Oracle Data Warehousing Guide. Overview of ETL. Release 2 (9.2) [en línia] Oracle Docs [Consultat: 29-Febrer-2016] Disponible en Internet: https://docs.oracle.com/cd/A97630_01/server.920/a96520/ettoverv.htm

[11] Oracle Data Warehousing Guide. Extraction in Data Warehouses. Release 2 (9.2) [en línia] Oracle Docs [Consultat: 29-Febrer-2016] Disponible en Internet: https://docs.oracle.com/cd/A97630_01/server.920/a96520/extract.htm

[14] Oracle Data Warehousing Guide. Loading and Transformation. Release 2 (9.2) [en línia] Oracle Docs [Consultat: 29-Febrer-2016] Disponible en Internet: https://docs.oracle.com/cd/A97630_01/server.920/a96520/transfor.htm

[15] Trello. [en línia] Disponible a Internet: https://trello.com/

[15] MongoDB Docs and Manual [en línia] Disponible a Internet: https://docs.mongodb.com/manual/

ANNEX A: DIAGRAMA DE GANTT

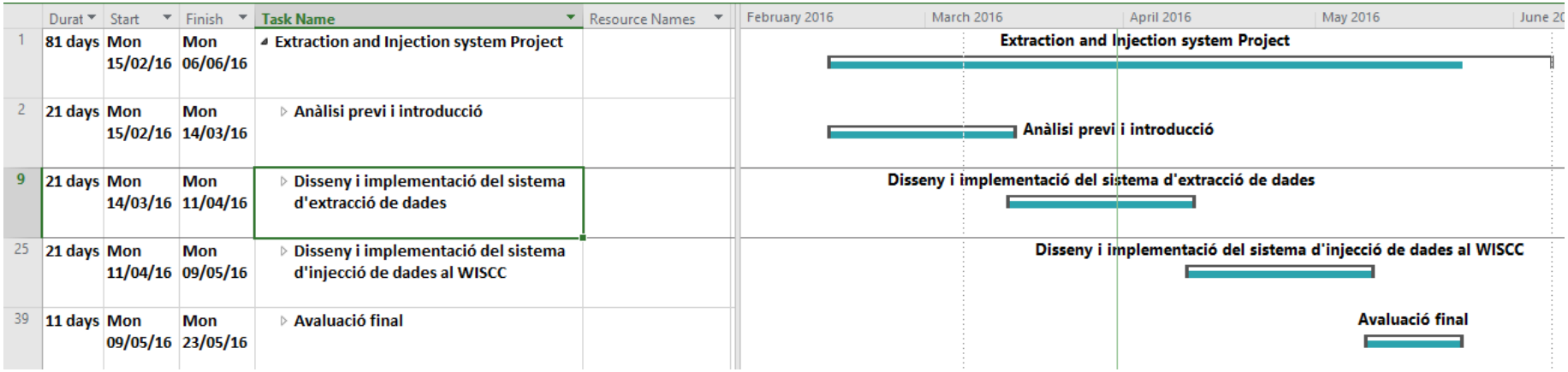

Figura 11 Visió global del projecte

	Durat	Start	Finish	Task Name	Resource Names
2	**21 days**	**Mon 15/02/16**	**Mon 14/03/16**	**Anàlisi previ i introducció**	
3	**21 days**	**Mon 15/02/16**	**Mon 14/03/16**	**Inception**	
4	3 days	Mon 15/02/16	Wed 17/02/16	Definició de l'abast del projecte	Cap de projecte
5	5 days	Mon 15/02/16	Fri 19/02/16	Estat de l'Art	Cap de projecte
6	15 days	Mon 22/02/16	Fri 11/03/16	Gestió del projecte. Planificació temporal i de costos	Cap de projecte
7	16 days	Mon 22/02/16	Mon 14/03/16	Anàlisi dels requisits generals del projecte	Analista
8	16 days	Mon 22/02/16	Mon 14/03/16	Implantació, introducció i adecuació a l'entorn de treball	Tots

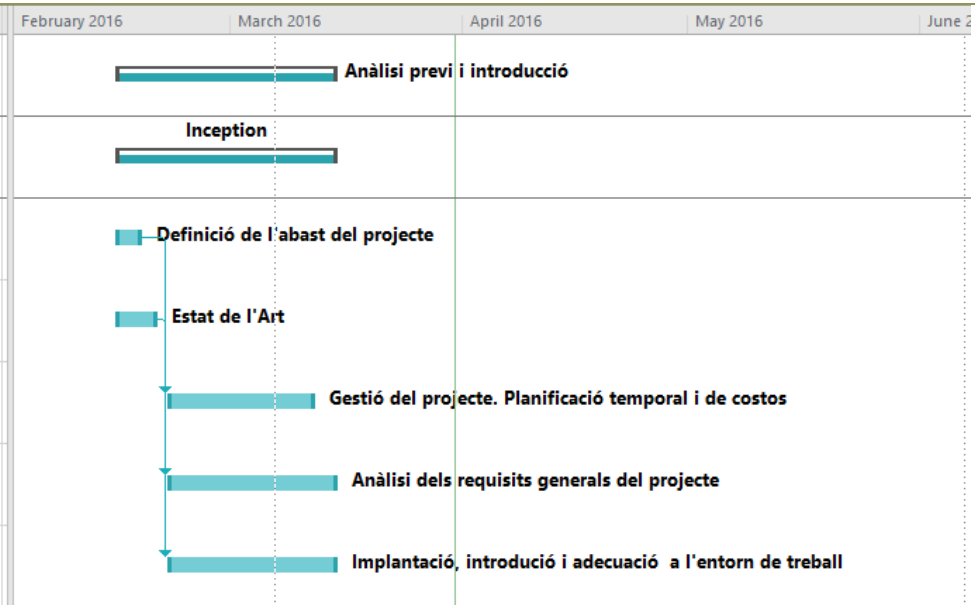

Figura 12 Anàlisi previ i introducció: Inception

	Durat	Start	Finish	Task Name	Resource Names
9	**21 days**	**Mon 14/03/16**	**Mon 11/04/16**	**Disseny i implementació del sistema d'extracció de dades**	
10	**11 days**	**Mon 14/03/16**	**Mon 28/03/16**	**Sprint 2**	
11	1 day	Mon 14/03/16	Mon 14/03/16	Planificació inicial Sprint	Tots
12	3 days	Mon 14/03/16	Wed 16/03/16	Anàlisi de requisits del sistema d'extracció	Analista
13	3 days	Mon 14/03/16	Wed 16/03/16	Anàlisi tècnic del sistema d'extracció	Arquitecte
14	5 days	Mon 14/03/16	Fri 18/03/16	Estudi de la API Web proporcionada per DHIS2	Programador i Arquitecte
15	3 days	Thu 17/03/16	Mon 21/03/16	Model conceptual del sistema d'extracció: Pipeline	Analista
16	4 days	Tue 22/03/16	Fri 25/03/16	Implementació del sistema funcional de flux de dades	Programador
17	2 days	Fri 25/03/16	Mon 28/03/16	Proves I validació Sprint	Responsable de proves
18	0 days	Mon 28/03/16	Mon 28/03/16	Entrega final Sprint 2	

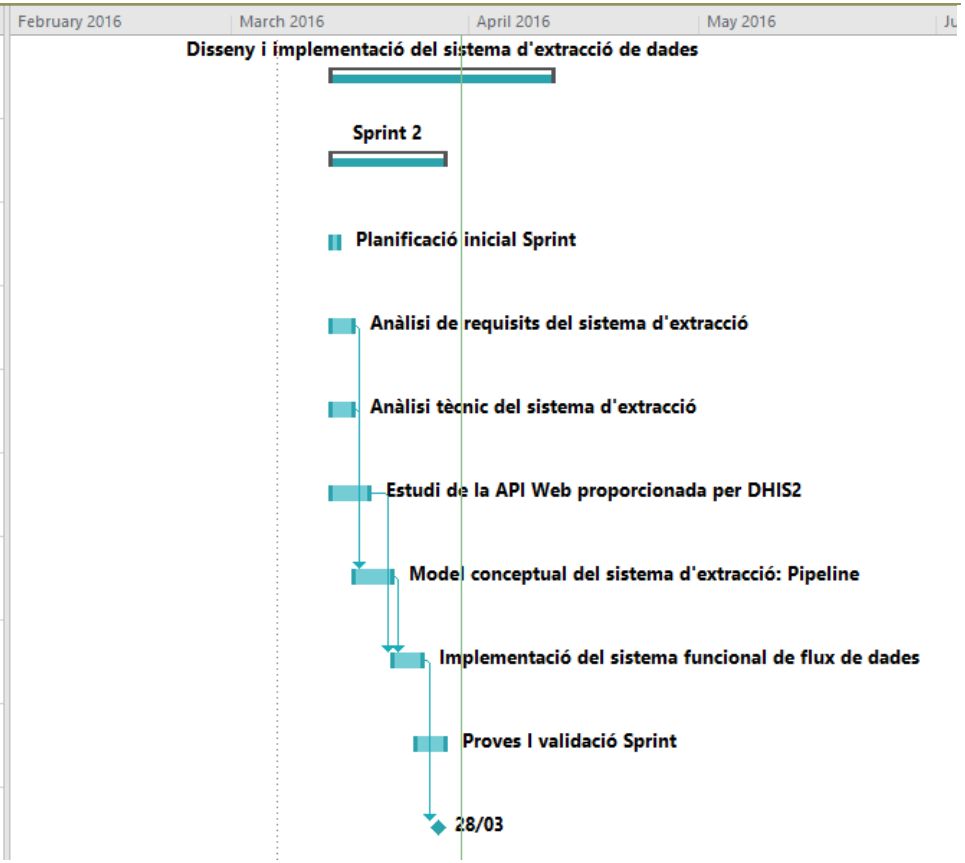

Figura 13 Disseny i implementació del sistema d'extracció de dades: Sprint 2

	Durat	Start	Finish	Task Name	Resource Names
9	**21 days**	**Mon 14/03/16**	**Mon 11/04/16**	**Disseny i implementació del sistema d'extracció de dades**	
10	**11 days**	**Mon 14/03/16**	**Mon 28/03/16**	**Sprint 2**	
19	**11 days**	**Mon 28/03/16**	**Mon 11/04/16**	**Sprint 3**	
20	1 day	Mon 28/03/16	Mon 28/03/16	Planificació inicial Sprint	Tots
21	10 days	Mon 28/03/16	Fri 08/04/16	Implementació dels fitxers I processos addicionals pel control de les execucions I dels fitxers d'anotació del historial.	Programador
22	8 days	Mon 28/03/16	Wed 06/04/16	Acoblament dels fitxers extrets de les instàncies Locals de DHIS2	Arquitecte i programador
23	2 days	Wed 06/04/16	Thu 07/04/16	Proves I validació Sprint	Responsable de proves
24	0 days	Mon 11/04/16	Mon 11/04/16	Entrega final Sprint 3	

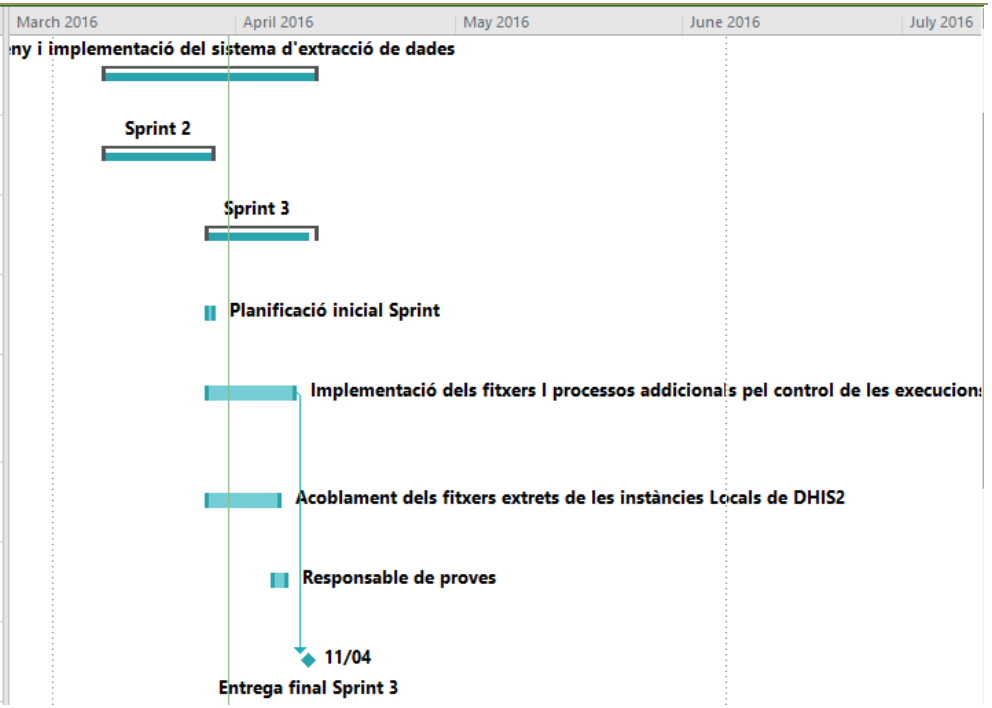

Figura 14 Disseny i implementació del sistema d'extracció de dades: Sprint 3

	Durat	Start	Finish	Task Name	Resource Names
25	**21 days**	**Mon 11/04/16**	**Mon 09/05/16**	**Disseny i implementació del sistema d'injecció de dades al WISCC**	
26	**11 days**	**Mon 11/04/16**	**Mon 25/04/16**	**Sprint 4**	
27	1 day	Mon 11/04/16	Mon 11/04/16	Planificació inicial Sprint	Tots
28	4 days	Mon 11/04/16	Thu 14/04/16	Disseny conceptual de la primera capa del repositori Data Lake del sistema WISCC	Analista
29	4 days	Mon 11/04/16	Thu 14/04/16	Disseny de la gestió de versions del registre de dades en el Data lake	Arquitecte
30	5 days	Fri 15/04/16	Thu 21/04/16	Implementació d'una versió primària d'injecció de dades al repositori	Programador
31	2 days	Thu 21/04/16	Fri 22/04/16	Proves I validació Sprint	Responsable de proves
32	0 days	Mon 25/04/16	Mon 25/04/16	Entrega final Sprint 4	

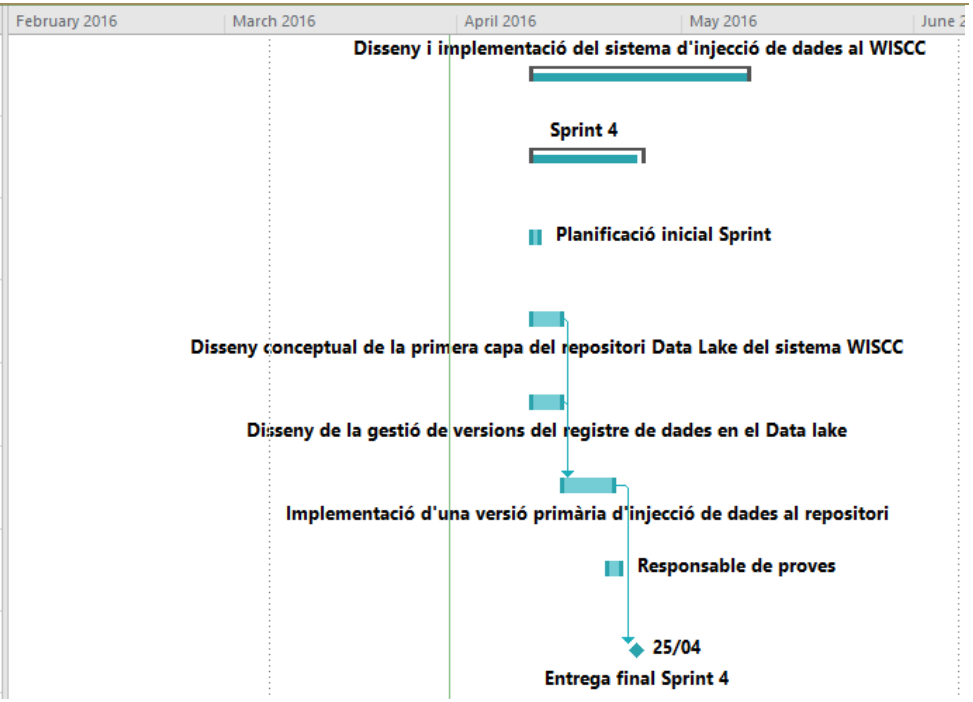

Figura 15 Disseny i implementació del sistema d'injecció de dades al WISCC: Sprint 4

	Durat	Start	Finish	Task Name	Resource Names
25	**21 days**	**Mon 11/04/16**	**Mon 09/05/16**	**Disseny i implementació del sistema d'injecció de dades al WISCC**	
26	**11 days**	**Mon 11/04/16**	**Mon 25/04/16**	**Sprint 4**	
33	**11 days**	**Mon 25/04/16**	**Mon 09/05/16**	**Sprint 5**	
34	1 day	Mon 25/04/16	Mon 25/04/16	Planificació inicial Sprint	Tots
35	5 days	Mon 25/04/16	Fri 29/04/16	Implementació completa de l'acoblament amb el sistema central de gestió de les metadades per la correcta formació dels fitxers	Programador
36	3 days	Mon 02/05/16	Wed 04/05/16	Implementació d'una versió final del sistema d'injecció amb la gestió de versions	Programador
37	2 days	Thu 05/05/16	Fri 06/05/16	Proves I validació Sprint	Responsable de proves
38	0 days	Mon 09/05/16	Mon 09/05/16	Entrega final Sprint 5	

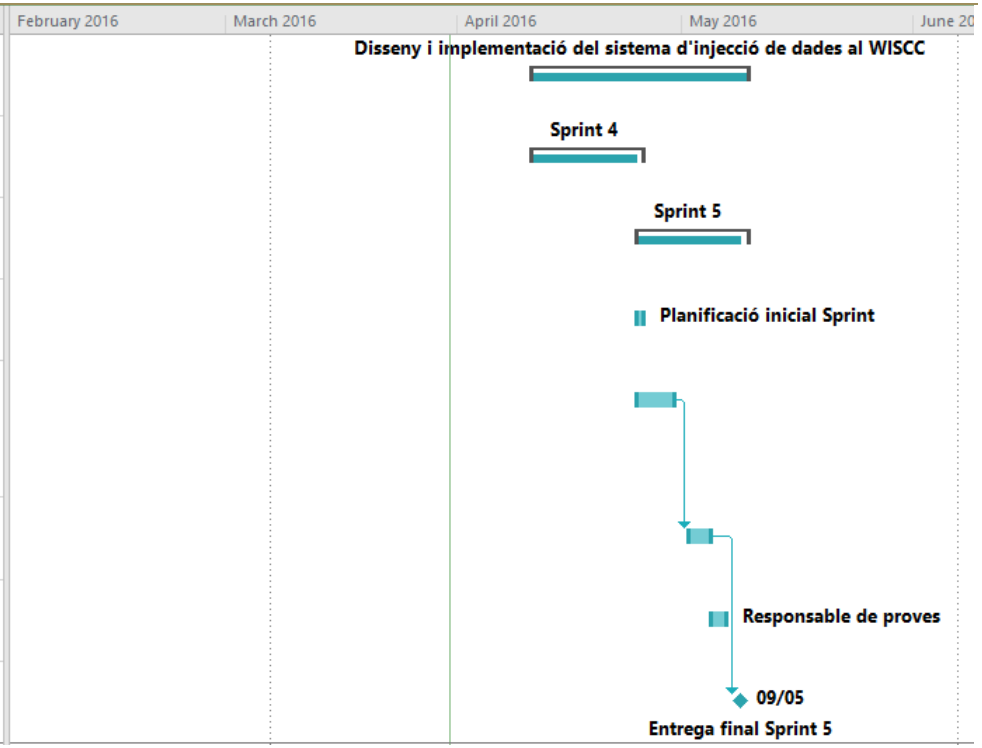

Figura 16 Disseny i implementació del sistema d'injecció de dades al WISCC: Sprint 5

	Durat	Start	Finish	Task Name	Resource Names
39	**11 days**	**Mon 09/05/16**	**Mon 23/05/16**	**Avaluació final**	
40	**11 days**	**Mon 09/05/16**	**Mon 23/05/16**	**Sprint 6**	
41	1 day	Mon 09/05/16	Mon 09/05/16	Planificació inicial Sprint	Tots
42	5 days	Mon 09/05/16	Fri 13/05/16	Instal·lació global en el sistema central del sistema	Arquitecte
43	4 days	Mon 16/05/16	Thu 19/05/16	Avaluació final	Tots
44	2 days	Thu 19/05/16	Fri 20/05/16	Proves I validació Sprint	Responsable de proves
45	0 days	Mon 23/05/16	Mon 23/05/16	Entrega Sprint 6	

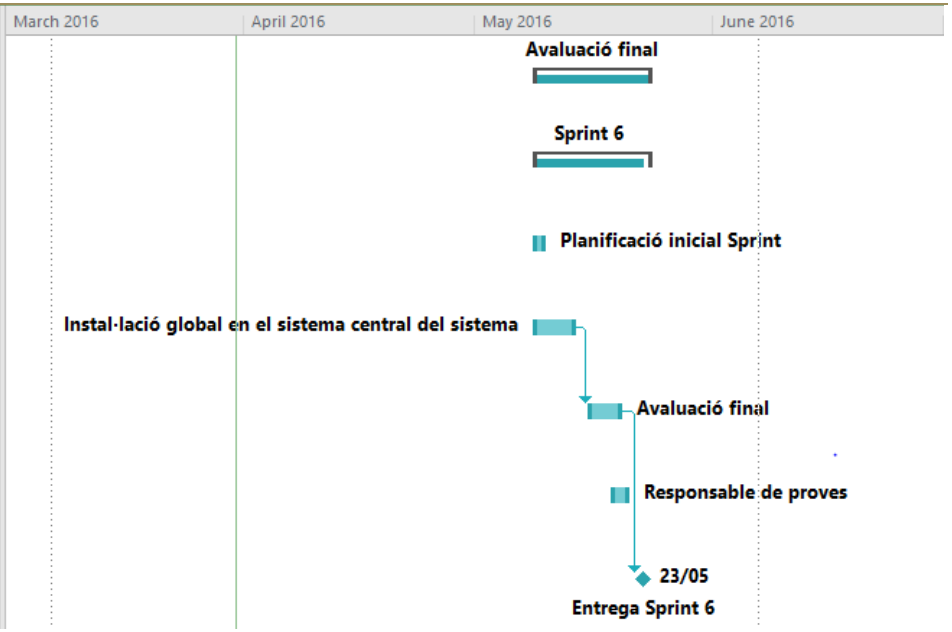

Figura 17 Avaluació final: Sprint 6

ANNEX B: DIAGRAMA DE GANTT DE DESPRÉS DE LES DESVIACIONS

Durat	Start	Finish
11 days	**Mon 11/04/16**	**Mon 25/04/16**
1 day	Mon 11/04/16	Mon 11/04/16
4 days	Mon 11/04/16	Thu 14/04/16
4 days	Mon 11/04/16	Thu 14/04/16
5 days	Fri 15/04/16	Thu 21/04/16
2 days	Thu 21/04/16	Fri 22/04/16
0 days	Mon 25/04/16	Mon 25/04/16

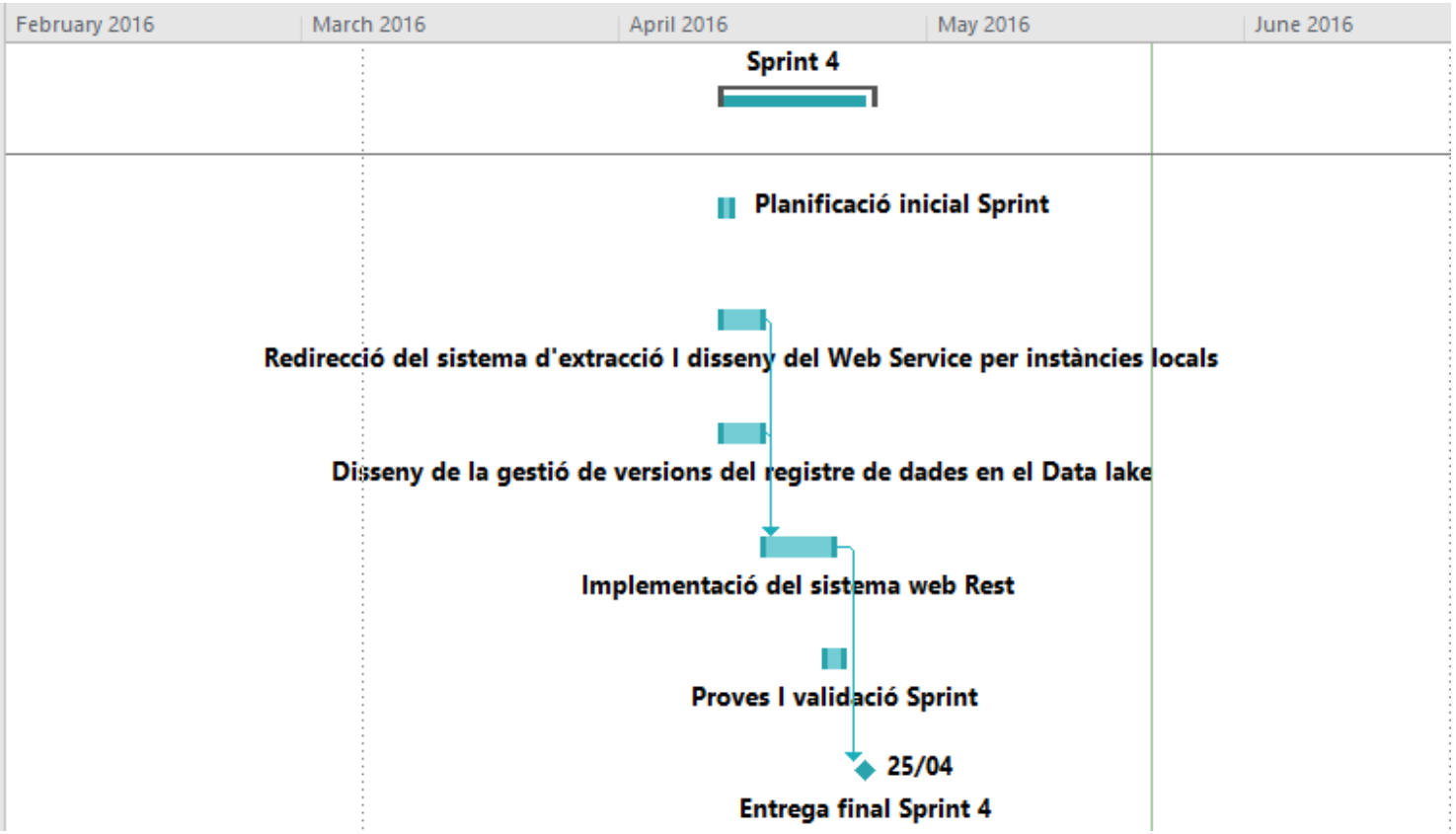

Figura 18 Desviacions de la planificació Sprint 4

Durat	Start	Finish
11 days	**Mon 25/04/16**	**Mon 09/05/16**
1 day	Mon 25/04/16	Mon 25/04/16
5 days	Mon 25/04/16	Fri 29/04/16
9 days	Tue 26/04/16	Fri 06/05/16
9 days	Tue 26/04/16	Fri 06/05/16
1 day	Fri 06/05/16	Fri 06/05/16
2 days	Fri 06/05/16	Mon 09/05/16
0 days	Mon 09/05/16	Mon 09/05/16

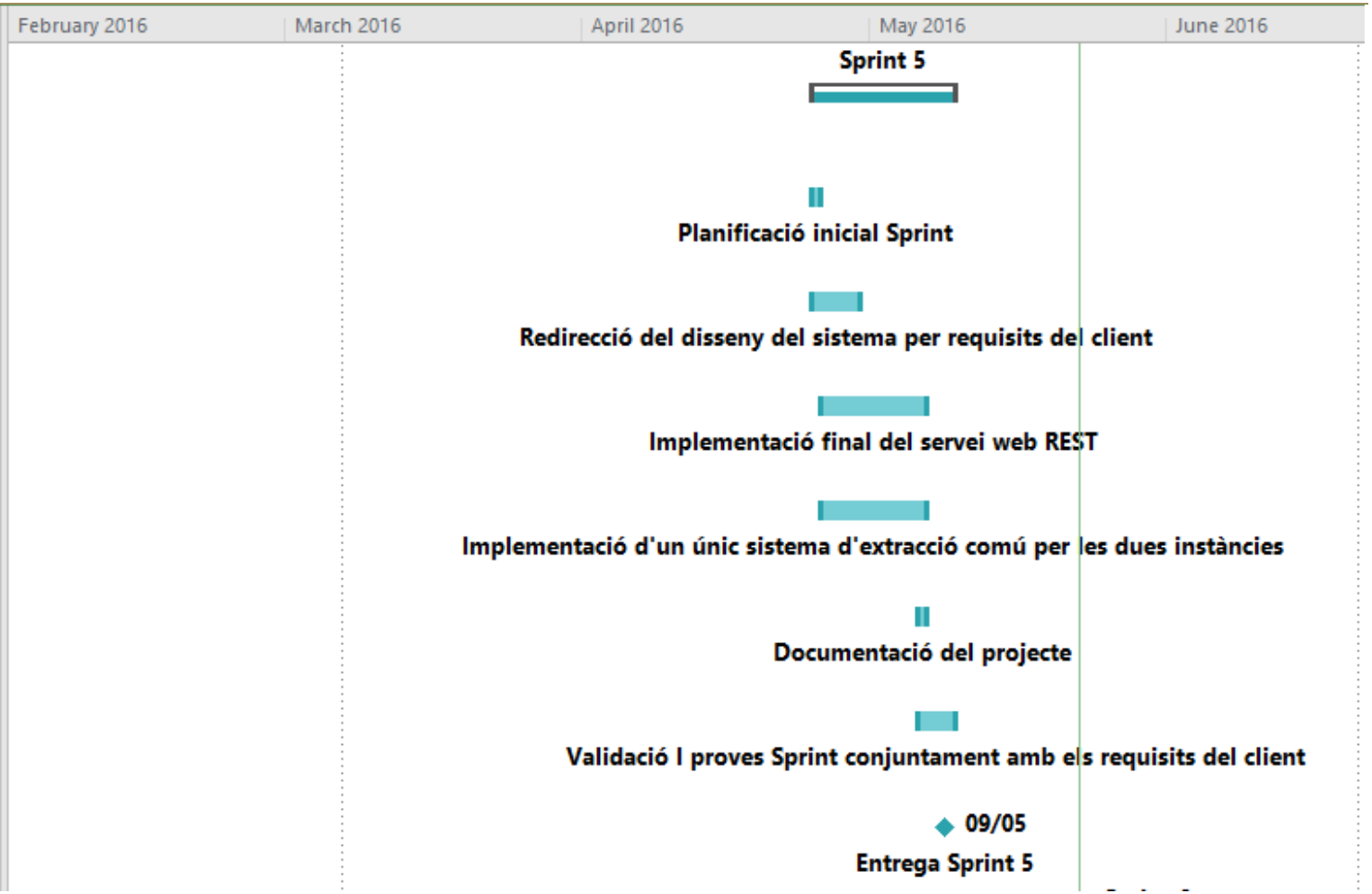

Figura 19 Desviacions de la planificació Sprint 5

Durat	Start	Finish
11 days	**Mon 23/05/16**	**Mon 06/06/16**
1 day	Mon 23/05/16	Mon 23/05/16
5 days	Mon 23/05/16	Fri 27/05/16
3 days	Mon 30/05/16	Wed 01/06/16
2 days	Tue 31/05/16	Wed 01/06/16
0 days	Mon 06/06/16	Mon 06/06/16

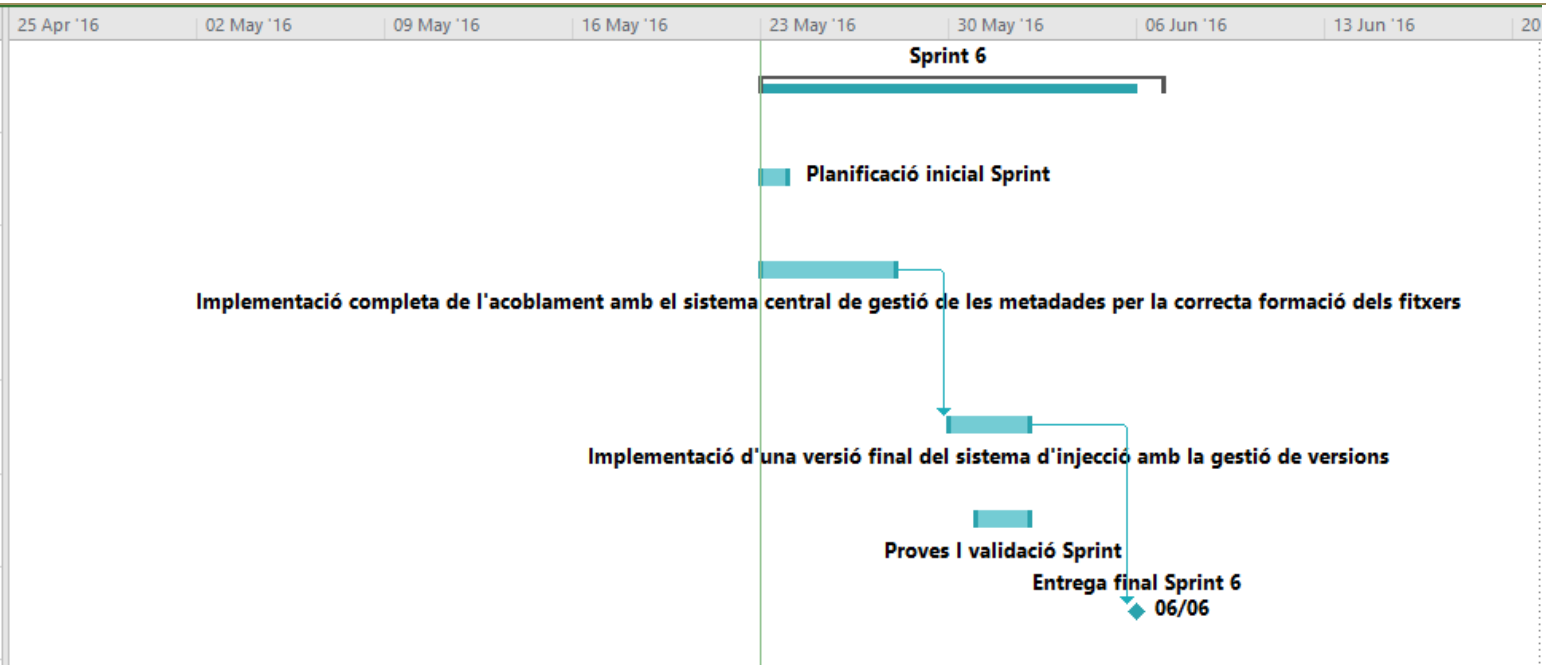

Figura 20 Desviacions de la planificació Sprint 6

Durat	Start	Finish
11 days	**Mon 06/06/16**	**Mon 20/06/16**
1 day	Mon 06/06/16	Mon 06/06/16
5 days	Mon 06/06/16	Fri 10/06/16
4 days	Mon 13/06/16	Thu 16/06/16
11 days	Mon 06/06/16	Sun 19/06/16
3 days	Thu 16/06/16	Sun 19/06/16
0 days	Sun 19/06/16	Sun 19/06/16

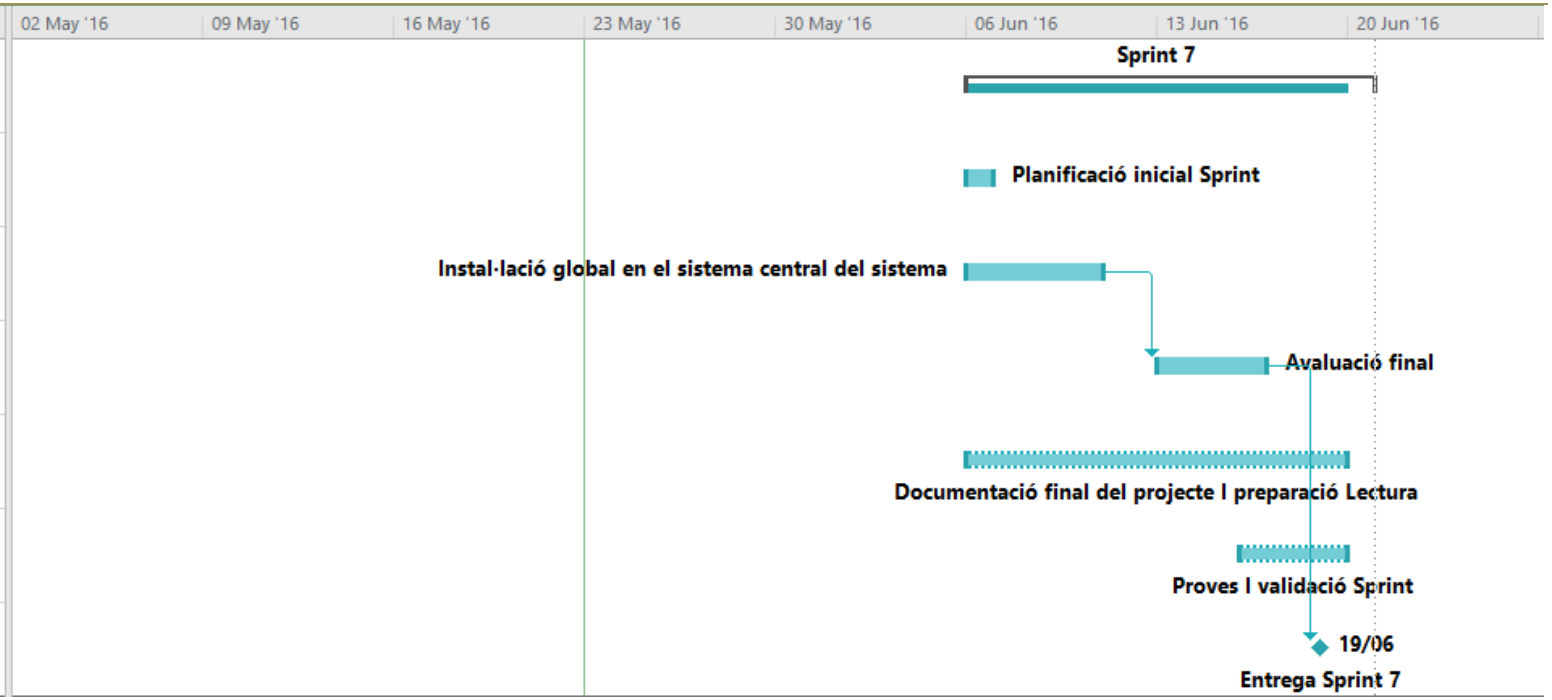

Figura 21 Desviacions de la planificació Sprint 7


FIB
UPC
UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH